IJCSIS Vol. 13 No. 5, May 2015
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA


Cornell University
Library


Cogprints


Google scholar


.docstoc
find and share professional documents


ScientificCommons

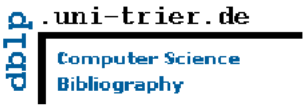

View my documents on
Scribd


BASE
Bielefeld Academic Search Engine


SCIRUS
search engine for science


SciRate.com


CiteSeerx beta


dblp .uni-trier.de
Computer Science
Bibliography


Q·Sensei


DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS


EBSCO
HOST


ProQuest

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 2015 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 Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.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 Managing Editor

It is a honor for the editorial board to present the May 2015 issue of International Journal of Computer Science and Information Security (IJCSIS). The purpose of this edition is to publicize experimental and theoretical research from both industry and academia in the broad areas of Computer Science, ICT & Security and further bring together people who work in the relevant areas. As the editors of this issue, we are glad to see variety of articles focusing on the major topics of innovation and computer science; computer security, interdisciplinary applications, information technologies etc. This journal promotes excellent research publications which offer significant contribution to the computer science knowledge and which are of high interest to a wide academic/research/practitioner audience. As a scholarly open access peer-reviewed journal, we provide an outlet for quality research & academic publications and support universal access for international scientific community to scientific knowledge; and the creation and dissemination of scientific and technical information.

IJCSIS archives all publications in major academic/scientific databases. Indexed by the following International agencies and institutions: Google Scholar, CiteSeerX, Cornell's University Library, Ei Compendex, Scopus, DBLP, DOAJ, ProQuest and EBSCO. Moreover, Google Scholar reported increased in number cited papers published in IJCSIS. We support researchers to succeed by providing high visibility & impact value, prestige and efficient publication process & service.

We thank our authors and look forward to your collaboration. As final words, Publish, Get Cited & Make An Impact.

For further questions 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. 13, No. 5, May 2015 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

TABLE OF CONTENTS

1. Paper 30011502: A Survey on Detection of Sinkhole Attack in Wireless Sensor Network (pp. 1-9)
George W. Kibirige & Camilius Sanga,, Department of Informatics, Sokoine University of Agriculture, SUA, Morogoro, Tanzania

Abstract - Wireless Sensor Network (WSN) consists of large number of low-cost, resource-constrained sensor nodes. The constraints of the wireless sensor node is their characteristics which include low memory, low computation power, they are deployed in hostile area and left unattended, small range of communication capability and low energy capabilities. Base on those characteristics makes this network vulnerable to several attacks, such as sinkhole attack. Sinkhole attack is a type of attack were compromised node tries to attract network traffic by advertise its fake routing update. One of the impacts of sinkhole attack is that, it can be used to launch other attacks like selective forwarding attack, acknowledge spoofing attack and drops or altered routing information. It can also used to send bogus information to base station. This paper is focus on exploring and analyzing the existing solutions which used to detect and identify sinkhole attack in wireless sensor network. The analysis is based on advantages and limitation of the proposed solutions.

Keywords: Wireless sensor network (WSN), sinkhole attack, detection of sinkhole attack

2. Paper 30041501: Decomposition and Denoise of an Image Using Singular Value Decomposition (pp. 10-12)
Viji Paul, Department of Information Technology, Salalah College of Technology, Sultanate of Oman
P. Selvaraju, Dept. of Computer Science & Engg, Vel Tech MultiTech Dr.RR and Dr.SR Engineering College, Avadi, Chennai, India
D. Nagarajan, Department of Information Technology, Salalah College of Technology, Sultanate of Oman

Abstract - Image decomposition is now essential for transmission and storage in database. Singular Value Decomposition is a decomposition technique for calculating the singular values, pseudo-inverse and rank of a matrix. The conventional way of doing this was to convert a matrix to row echelon form. The rank of a matrix is then given by the number of nonzero rows or columns of the echelon form. Singular value decomposition is one of the methods to compress and denoise the images. The main focus of this paper is to decompose and denoise an image using singular value decomposition.

Keywords: SVD, decomposition, denoise.

3. Paper 30041503: An Improvement on Fragmentation in Distribution Database Design Based on Knowledge-Oriented Clustering Techniques (pp. 13-17)
Van Nghia Luong, Department of Information Technology, Pham Van Dong University, Quang Ngai, Viet Nam
Ha Huy Cuong Nguyen, Department of Information Technology, Quang Nam University, Quang Nam, Viet Nam
Van Son Le, Da Nang University of Education, Da Nang University, Da Nang, Viet Nam

Abstract — The problem of optimizing distributed database includes: fragmentation and positioning data. Several different approaches and algorithms have been proposed to solve this problem. In this paper, we propose an algorithm that builds the initial equivalence relation based on the distance threshold. This threshold is also based on knowledge- oriented clustering techniques for both of horizontal and vertical fragmentation. Similarity measures used in the algorithms are the measures developed from the classical measures. Experimental results carrying on the small data set match fragmented results based on the classical algorithm. Execution time and data fragmentation significantly reduced while the complexity of our algorithm in the general case is stable.

Keywords — Vertical Fragmentation; Horizontal Fragmentation; Similarity Measure; Clustering Techniques; knowledge-oriented clustering techniques.

4. Paper 30041505: Visualizing Object-oriented Software for Understanding and Documentation (pp. 18-27)
Ra'Fat AL-msie'deen, Department of Information Technology, Mutah University, Al-Karak, Jordan

Abstract — Understanding or comprehending source code is one of the core activities of software engineering. Understanding object-oriented source code is essential and required when a programmer maintains, migrates, reuses, documents or enhances source code. The source code that is not comprehended cannot be changed. The comprehension of object-oriented source code is a difficult problem solving process. In order to document object-oriented software system there are needs to understand its source code. To do so, it is necessary to mine source code dependencies in addition to quantitative information in source code such as the number of classes. This paper proposes an automatic approach, which aims to document object-oriented software by visualizing its source code. The design of the object-oriented source code and its main characteristics are represented in the visualization. Package content, class information, relationships between classes, dependencies between methods and software metrics is displayed. The extracted views are very helpful to understand and document the object-oriented software. The novelty of this approach is the exploiting of code dependencies and quantitative information in source code to document object-oriented software efficiently by means of a set of graphs. To validate the approach, it has been applied to several case studies. The results of this evaluation showed that most of the object-oriented software systems have been documented correctly.

Keywords- Vsound; software engineering; software documentation; software visualization; software understanding; software maintenance; software evolution; software reuse; change impact analysis; object-oriented source code; reverse engineering.

5. Paper 30041508: The Knowledge Management Strategies Used as a Tool within and through Strategic Consulting Firms to Increase the Organisational Performance (pp. 28-33)
Fatimetou Zahra Mohamed Mahmoud, Ahmed Mohamed Mahmoud, Faculty of Information and Communication Technology, IIUM, Kuala Lumpur, Malaysia
Jamaludin Ibrahim, Faculty of Information and Communication Technology, IIUM, Kuala Lumpur, Malaysia

Abstract - As a matter of fact, the rising awareness in organisations to not focus only in the physical assets in the organisation but also in other paramount assets such as knowledge was the launching point for organisations to go to the implementation of knowledge management and creating knowledge management strategies codified or personalized in order to overcome issues and plan for its future sustainability and development. Actually, the implementation of knowledge management strategy is the strategic consulting company internal asset and external mean to get value. Although, after the implementation of knowledge management strategy in some companies they face the problem of the low revenue and efficiency of the strategy implemented. That's why the combination of the codification and personalization make the organisation able to create, share, save and re-use knowledge and exchange it in the organization using technology or informal communication networks in order to raise the innovation in the work environment and the creation of new ideas and services. Also, this knowledge stored and codified especially about the organization customers can be considered as the basic knowledge to take the strategic decisions and by KM the organisation will be able to extract, understand and well use this knowledge in order to strengthen the relation with its customers and allow them to become the organisation's partners and advisors. Indeed, that's mean that the organization strategy must be renewable and in corporation with the overall business model and focus on the core competence in the organisation and the management of the customer knowledge and interactions, this will influence positively and upgrade the organisational performance in any organisation.

Keywords: Knowledge management, Strategy, Consulting firms, organisational performance.

6. Paper 30041510: Discrete Fourier Transform Attack on WG-7 Cipher (pp. 34-38)
Rashid Zulfiqar & Mehreen Afzal, College of Telecommunication(MCS), National University of Sciences and Technology, Islamabad, Pakistan

Abstract — This paper presents application of Discrete Fourier Transform(DFT) attack on Stream Cipher Welch Gong(WG)-7. WG-7 is a lightweight, hardware oriented stream cipher that uses a word oriented Linear feedback shift register (LFSR) and a nonlinear WG transformation that acts on the LFSR output word. The cipher has been designed to work in a resource constrained environment by Yiyuan Luo, Qi Chai, Guang Gong and Xuejia Lai as

variant of original WG Cipher. This paper aims at faster recovery of keystreams than predicted complexity of the DFT attack by the designers. The proposed DFT attack recovers the keystream with a time complexity of at the most 2^22 and keybits complexity of at the most 2^19 by employing the annihilator in the structure of the cipher. The proposed attack is even efficient than Algebraic attack and Fast Algebraic attack on the cipher.

Index Terms—WG-7, Discrete Fourier Transform Attack, Key Recovery Attack, Discrete Fourier Spectra Attack.

7. Paper 30041511: A Comparison of GPU based Networking Algorithms (pp. 39-47)

Jaideep Patel & Nilay Khare, Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal, India

Abstract — Modern high performance routers rely on customized hardware solutions, therefore difficult to adapt to continuously changing network protocols. While software routers provide the flexibility and programmability, thereby achieving a lower throughput. Modern GPUs offer significant computing power, and its data parallel computing matches the packet processing on routers. As the routing table size increases along with the physical link speeds, IP address lookup has been a difficult problem at routers. It has been a challenge to achieve a cost effective high-performance IP lookup. With the increasing Internet traffic and fast changing network protocols, the next generation routers have to satisfy the need for throughput, scalability, flexibility and QoS. The aim of this paper is to compare different GPU based networking algorithm on different performance metrics and to discuss their benefits and drawbacks.

Keywords-Networking algorithms; Software Router; GPU; IPLookup; QoS.

8. Paper 30041514: A Survey-Defect Detection and Classification for Fabric Texture Defects in Textile Industry (pp. 48-56)

Ms. Shweta Loonkar & Dr. Dhirendra Mishra, Computer Engineering Department, MPSTME, NMIMS University, Mumbai-56

Abstract- Textile industry is one of the largest and oldest sectors in the India and has a formidable presence in national economy in terms of output, investment and employment. Due to increasing demand for quality fabrics it is thus important to produce the defect free high quality fabric. Visual inspection system consumes a lot of time and are error prone. The price of the fabric is reduced to 45%-65% due to presence of various defects. The purpose of this paper is to automate the detection and classification of texture defects by computerize software. The proposed method uses a statistical based approach for the inspection and detection of the defect on woven/knitted fabric collected from the textile industry. In this the images are acquired, pre-processed, restored and normalized to extract the statistical feature using computer vision. The extracted features are given an input to the artificial neural network decision tree classifier to compute the weighted factor for detecting and classifying the type of defects. An automatic defect detection system can increase the texture defect detection percentage and will reduce the fabrication and labour cost and improves the quality of the product.

Keywords: Defect detection, Statistical approach, Computer vision, Decision tree classifier, neural network.

9. Paper 30041516: Management System of Children with Autism (pp. 57-66)

Elham Mohammed Thabit AbdAlameer, Karbala University, College of Science, Department of Computer Science, Iraq, Karbala

Abstract- Autism is one of the most prevalent chronic disease; it is a developmental disability that causes problems with communication and social interaction shared with repetitive demeanour. Some behaviors could occur before age three and appear through delays in diverse skills which develop from childhood to adulthood, even though behaviors may optimize over time. The disorder includes a large spectrum of symptoms, levels of impairment, and skills. It varies in severity from a handicap which somewhat limits an otherwise normal life to a devastating disability which may need institutional care. However, optimization of health care is probable to have a good effect on progress of disease, quality of life, and functional outcome. This paper will design and built a reliable system based Electronic healthcare technologies to care with autistic child. It is centralized database which contains the patients profiles; the system will enable the nurse to access the database and enter details of patients for subject them to a set of tests; then these data will be held on a database and the system will identify the type of autism to

determine the suitable treatment to each type. This proposed clinical system will serve the patient and health care providers by reducing healthcare costs and enable parents to help their autistic children by using the best way to treat them.

Keywords- ASDs, Autistic children, Neurobehavioral, Clinical system, Ehealth care systems, Pediatricians.

10. Paper 30041520: Requirement Tracing using Term Extraction (pp. 67-72)

Najla Al-Saati, Software Engineering Dept, College of Computer Sciences & Mathematics, Mosul, Iraq
Raghda Abdul-Jaleel, Software Engineering Dept, College of Computer Sciences & Mathematics, Mosul, Iraq

Abstract - Requirements traceability is an essential step in ensuring the quality of software during the early stages of its development life cycle. Requirements tracing usually consists of document parsing, candidate link generation and evaluation and traceability analysis. This paper demonstrates the applicability of Statistical Term Extraction metrics to generate candidate links. It is applied and validated using two datasets and four types of filters two for each dataset, 0.2 and 0.25 for MODIS, 0 and 0.05 for CM1. This method generates requirements traceability matrices between textual requirements artifacts (such as high-level requirements traced to low-level requirements). The proposed method includes ten word frequency metrics divided into three main groups for calculating the frequency of terms. The results show that the proposed method gives better result when compared with the traditional TF-IDF method.

Keywords - Requirements Traceability; Traceability Analysis; Candidate Link Generation; Parsing; Term Extraction; Word Frequency Metrics.

11. Paper 30041521: A Fast and Accurate Pupil and Iris Localization Method Usable with a Regular Camera (pp. 73-82)

Gheis Mohammadi, Department of Computer Engineering, Science and Research branch, Islamic Azad University of Tehran, Iran
Jamshid Shanbehzadeh, Department of Computer Engineering, Kharazmi University, Tehran, Iran

Abstract — The fundamental step in implementing an eye-based interface is the exact localization of user eye pupil and iris and separating them from other parts of the eye. This paper presents a fast adaptive iris localization scheme that can be used by low resolution cameras or webcams. This method finds the pupillary and iris boundaries accurately with low computational cost under variable and noisy lighting conditions. Based on the fact that the pixels of the pupil are darker than other regions, this method locates its boundary circle and then uses it to localize the outer boundary of the iris. The first step to identify the inner boundary is to calculate a threshold value to separate the pupil's pixels from other parts of the eye. Next, we search in the neighbourhood of the initial threshold value for the most appropriate threshold value. This value binaries image and Circular Hough Transform detects circles with certain values. After remove most circles, the most accurate one is chosen. This circle is the boundary of the pupil. Finally, the outer boundary of the iris is identified. The performance of the proposed algorithm was assessed by using it to segment both low and high resolution images. We ran the experiments on UBIRIS v1.0 also, for comparison with other reported accuracies. The experimental results show that the proposed method has a 14ms detection time and 97.76% detection accuracy with an overall accuracy of 99.025%. These results can be achieved in low resolution images also that show an improvement in pupil and iris localizing performance in comparison with well-known methods.

Keywords: Eye localization; Eye tracking; Image analysis; Pupil and iris boundary; Robust Adaptive localization.

12. Paper 30041522: Preserving Trajectory Privacy in Participatory Sensing Applications (pp. 83-90)

Gauri R Virkar, Sanchika A Bajpai, Department Of Computer Engineering, BSIOER, Wagholi, Pune, India

Abstract — With the advancement of technology in fields of wireless communication, different mobile communicating devices equipped with variety of embedded sensors and powerful sensing have been emerged. Participatory sensing is the process that enables individuals to collect, analyze and share local knowledge with their own mobile devices. Although the use of participatory sensing offers numerous benefits on deployment costs, availability, spatial- temporal coverage, energy consumption and so forth, it has certain threats which may be compromise the participator’s location and their trajectory data. Henceforth, to ensure the participators’ privacy is

the most urgent task. The existing proposals emphasized more on participators' location privacy and very few of them consider the privacy of the trajectories. The theoretical mix zones model are been improved by considering time factor from the viewpoint of the graph theory and mix zone graph model has been presented. This model considers only sensitive trajectories for providing privacy thereby reducing overall information loss and storage space. Further, instead of defining single mix zone graph model, multiple mix zones are created in order to enhance the privacy of the participator's trajectories.

Keywords- Location privacy, mix zone graph model, multiple mix zone, participatory sensing, trajectories.

13. Paper 31031503: Security Analytics in Big Data Infrastructures (pp. 91-95)

K. Boukri & H.Chaoui, Systems Engineering Laboratory, the ADSI team, National School of Applied Sciences in Kenitra. Campus Universitaire, BP 241 Kenitra Morocco

Abstract - Big Data is usually so large and complex, and it has become an emerging hot topic in network security fields. How to deal with lots of safety data which are produced by heterogeneous network security devices, how to analyze and coordinate such Big Data network events must be studied to date [1-3]. This paper develops a kind of Big Data Analytics Security framework using collection, analysis and integration, event-correlation and scenario analysis technique to process the raw data gathered from Big Data infrastructure (BDI).

Keywords: Big Data Security, Big Data analytics, Intelligence Analysis, Intrusion detection

14. Paper 31051518: Switching of code between the users for enhancing the Security in OCDMA (pp. 96-100)

Sumit Gupta & Aditya Goel, Electronics and Communication Engineering Department, Maulana Azad National Institute of Technology, Bhopal, India

Abstract- A new technique is proposed for optical code division multiple accesses for enhancing the security against the eavesdropper. Switching of the pulse spectrum of code is performed between the two users and switching position of the pulse spectrum of code varies from group to group. In this technique every pulse of a code does not have direct information. Code detection probability of individual user decreases against the eavesdropper. The analysis and simulation result compares with the exiting method MQC, RD and MDW code.

Keywords: Random Diagonal (RD) Code, Modified Double Weight (MDW) Code, Zero cross correlation code (ZCC), Spectral Amplitude Coding (SAC)

A Survey on Detection of Sinkhole Attack in Wireless Sensor Network

George W. Kibirige
Department of Informatics
Sokoine University of Agriculture, SUA
Morogoro, Tanzania

Camilius Sanga,
Department of Informatics
Sokoine University of Agricultur, SUA
Morogoro, Tanzania

Abstract
Wireless Sensor Network (WSN) consists of large number of low-cost, resource-constrained sensor nodes. The constraints of the wireless sensor node is their characteristics which include low memory, low computation power, they are deployed in hostile area and left unattended, small range of communication capability and low energy capabilities. Base on those characteristics makes this network vulnerable to several attacks, such as sinkhole attack. Sinkhole attack is a type of attack were compromised node tries to attract network traffic by advertise its fake routing update. One of the impacts of sinkhole attack is that, it can be used to launch other attacks like selective forwarding attack, acknowledge spoofing attack and drops or altered routing information. It can also used to send bogus information to base station. This paper is focus on exploring and analyzing the existing solutions which used to detect and identify sinkhole attack in wireless sensor network. The analysis is based on advantages and limitation of the proposed solutions.

Keywords: Wireless sensor network (WSN), sinkhole attack, detection of sinkhole attack

I. INTRODUCTION

Wireless sensor network consists of small nodes with ability to sense and send data to base station [5]. Wireless sensor network is used in different applications example in military activities, which used to track movement of their enemy. It also used in fire detection and in healthy service for monitoring heart beat [2, 17, 13]. Unfortunately most of wireless network are deployed in unfriendly area and normally left unattended. Also most of their routing protocols do not consider security aspect due to resource constraints which include low computational power, low memory, low power supply and low communication range [8,9]. This constraint creates chance for several attackers to easily attack wireless sensor network. An example of attack is sinkhole attack. Sinkhole attack is implemented in network layer where an adversary tries to attract many traffic with the aim to prevent base station from receiving a complete sensing data from nodes [20].The adversary normally compromises the node and that node will be used to launch an attack. The compromised node send fake information to neighboring nodes about its link quality which used in routing metric to select best route during data transmission. Then all the packets from his neighbors pass through him before reach to base station. [22]. Sinkhole attack prevents base station from acquiring a complete and correct sensing data from nodes.

The purpose of this paper is to study existing solutions used to detect sinkhole attack. Different solutions which were used to detect and identified sinkhole attack were suggested by different researchers, such as Krontiris [14], Ngai et al [18] and Sheela et al [25]. Rule based detection solution were proposed by Krontiris et al[15] to detect sinkhole attack. All the rules were focused on node impersonation and were implanted in intrusion detection system. Then intruder was easily detected when they violate either of the rules. Another centralized solution which involve base station in detection process proposed by Ngai et al [18] A non cryptography scheme which used mobile agent in the network to prevent sinkhole attack was also proposed by Sheela et al [25]

The remainder of this paper is organized as follow. Section 2 discusses sinkhole attack and their attack mechanism in two different protocols. Section 3 presents the challenges in detection of sinkhole attack in wireless sensor network. Section 4 presents different approaches that proposed by different researchers to detect sinkhole attack. Finally, section 5 conclude this paper and proposed some future works.

II. SINKHOLE ATTACK

Sinkhole attack is an insider attack were an intruder compromise a node inside the network and launches an attack. Then the compromise node try to attract all the traffic from neighbor nodes based on the routing metric that used in routing protocol. When it managed to achieve that, it will launch an attack. Due to communication pattern of wireless sensor network of many to one communication where each node send data to base station, makes this WSN vulnerable to sinkhole attack (Ngai et al [18]).

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

The following subsections discuss the techniques use in MintRoute protocol and AODV protocol in launching sinkhole attack.

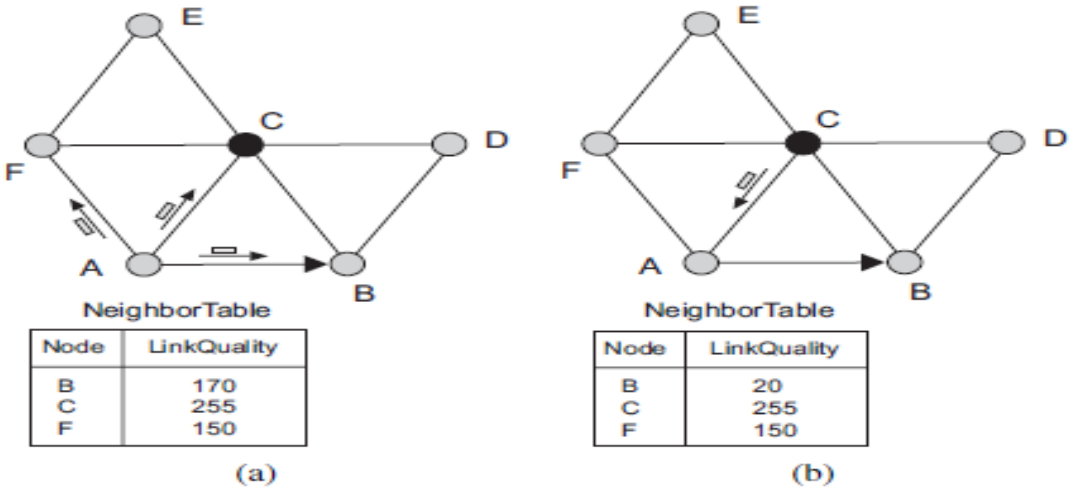


Figure 1: Sinkhole attack in MintRoute protocol (Krontiris, I[15])

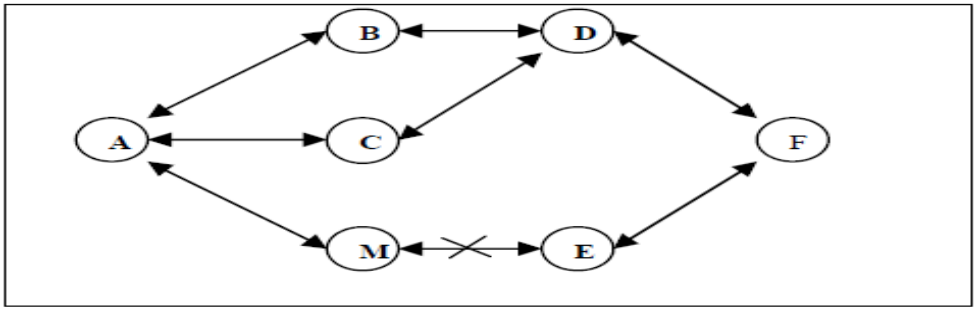


Figure 2: Sinkhole in TinyAODV protocol (Teng and Zhang,[27])

Sinkhole Attack in MintRoute Protocol

MintRoute protocol is a type of protocol which is commonly used in wireless sensor network. It was designed purposely for the wireless sensor network, it is light and suitable for sensor nodes which have minimum storage capacity, low computation power and limited power supply. MintRoute protocol uses link quality as a metric to choose the best route to send packet to the Base Station (Krontiris et al [15]).

Fig.1 shows six sensor nodes A, B, C, D, E, and F. Node C is malicious, and it is going to launch a sinkhole attack. The Figure 1(a) shows a route table of node A with IDs of its neighbors with their corresponding link quality. Originally the parent node was node B but node C advertises its link quality with a value of 255 which is maximum value. Node A is not going to change its parent node until the node B's link quality fall to 25 below the absolute value.

In Fig.1(b) the malicious node is sending new update route packet that the link quality fall up to 20 and impersonate node B so that node A believe the packet come from node B. Node A will update its route table and change the parent node to node C (Krontiris et al [15]). The attacker uses node impersonation to launch an attack.

Sinkhole Attack in TinyAODV Protocol

This is another explanation of sinkhole attack in wireless sensor network and this time the attack is launched under TinyAODV (Ad-hoc On Demand Vector) protocol. TinyAODV protocol is the same as AODV in MANET but this one is lighter compared to AODV and it was modified purposely for wireless sensor network [27]. The number of hops to base station is the routing metric that used in this protocol. Generally the route from source to destination is created when one of the nodes send a request, the source node sends a RREQ (Route request) packet to his neighbors when wants to send packet. Next one of the neighbors close to destination is reply by sending back RREP (Route Reply) packet, if not the packet is forwarded to other nodes close to that

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

destination. Finally, the source receives RREP packet from neighbor then select one node with less number of hops to destination.

The sinkhole node or compromised node launches an attack by send back RREP packet. In RREP packet it gives small number of hops which indicates close proximity to the base station. Then the source node decides to forward packet to sinkhole node. The compromised node then performs the same technique to its entire neighbors and tries to attract as much traffic as possible [27].

For instance, Fig.2 shows node M launches sinkhole attack in Tiny AODV. Node A sends RREQ to nodes BCM. However node M instead of broadcast to node E like nodes B and C does to node D, he replies back RREP to node A. Then node A will reject node B and C, then forward packet to M because node A and B are very far to F compare to node M.

III. CHALLENGES IN DETECTION OF SINKHOLE ATTACK IN WSNs

Based on the literature review of sinkhole attack in wireless sensor network, the following are the main challenges in detecting sinkhole attack in wireless sensor network

A. Communication Pattern in WSN;

All the messages from sensor nodes in wireless sensor network are destined to base station. This created opportunity for sinkhole to launch an attack. Sinkhole attacks normally occur when compromised node send fake routing information to other nodes in the network with aim of attracting as many traffic as possible. Based on that communication pattern the intruder will only compromised the nodes which are close to base station instead of targeting all nodes in the network. This is considered as challenges because the communication pattern itself provides opportunity for attack.

B. Sinkhole attack is unpredictable*;*

In wireless sensor network the packet are transmitted based on routing metric that used by different routing protocols [26]. The compromised node used its routing metric that used by routing protocol to lie to his neighbors in order to launch sinkhole attack. Then all the data from his neighbors to base station will pass through compromised node. For example the techniques used by compromised node in network that used TinyAODV protocol is different to the one used another protocol like MintRoute protocol. In MintRoute they used link quality as route metric while in Tiny AODV they used number of hop to base station as routing metric. Therefore the sinkhole attack techniques is changed based on routing metric of routing protocol

C. Insider Attack

Insider attack and outsider attack are two categories of attack in wireless sensor network. Outside attack is when intruder is not part of network. In inside attack the intruder compromises one of the legitimate node through node tempering or through weakness in its system software then compromised node inject false information in network after listen to secret information. Inside attack can disrupt the network by modifying routing packet. Through compromised node sinkhole attack attract nearly all the traffic from particular area after making that compromised node attractive to other nodes. The fact is that compromised node possesses adequate access privilege in the network and has knowledge pertaining to valuable information about the network topology this created challenges in detecting. Base to that situation even cryptographic cannot defend against insider attack although it provides integrity, confidentiality and authentication (Pathan, K [22]). Therefore the internal attack has more serious impact on victim system compared to outsider attack.

D. Resource Constraints;

The limited power supply, low communication range, low memory capacity and low computational power are the main constrained in wireless sensor network that hinder implementation of strong security mechanism. For example the strong cryptographic method that used in other network cannot be implemented in this network due to low computational power and low memory capacity. Therefore less strong key are considered which is compatible with available resources.

E. Physical attack;

A wireless sensor network normally deployed in hostile environment and left unattended. This provides a opportunity for an intruder to attack a node physically and get access to all necessary information [12].

IV. EXISTING APPROACHES

Many researchers have been working on wireless sensor field to provide security mechanism to suits the resource constrained due to growing demand of applications in sensitive areas. The following are the identified approaches that used by different researchers to detect and identified sinkhole attack in wireless sensor network. Those approaches are classified into rules based, key management, anomaly based, statistical method and hybrid based. The subsequent subsections described each of these categories and give examples of existing work that used that approach.

A. **Rule based**

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

The rules are designed based on the behavior or technique used to launch sinkhole attack. Then those rules are imbedding in intrusion detection system which runs on each sensor nodes. Those rules were then applied to the packet transmitted through the network nodes. If any node violates the rules is considered as adversary and isolated from the network.

Among the existing work which used rules based approach include Krontiris et al [14]. Krontiris used rule based approach to detect sinkhole attack. They create two rules and implanted in Intrusion detection system (IDS). When one of the rules is violated by one of the nodes, the intrusion detection system triggered an alarm but it does not provide node ID of compromised node. The first rule "for each overhead route update packet the ID of the sender must be different your node ID". The second rule "for each overhead route update packet the ID of the sender must be one of the node ID in your neighbors". Also Krontiris et al [15] used the same approaches. There are two rules, the first rule "rule for each overhead route update packet the ID of the sender must be one of node ID in your neighbors". The second rule "for each pair of parent and child node their link quality they advertise for the link between them, the difference cannot exceed 50.

B. Anomaly-based detection

In anomaly based detection the normal user behavior is defined and intrusion detection is searching for anything that is anomalous in the network. In this method intrusion is considered as anomalous activity because it looks abnormal compare to normal behavior. The rule based and statistical approaches are also included under anomaly based detection approach.

Tumrongwittayapak and Varakulsiripunth [29] proposed system that used RSSI (Received Signal Strength Indicator) value with the help of EM (Extra Monitor) nodes to detect sinkhole attack. The EM had high communication range and one of their functions is to calculate RSSI of node and send to base station with ID of source and next hop. This process happens instantly when node are deployed. Base station uses that RSSI value to calculate VGM (visual geographical map). That VGM shows the position of each node, then later when EM send updated RSSI value and base station identify there is change in packet flow from previous data this indicate there is sinkhole attack. The compromised node is identified and isolated from the network by base station using VGM value. However, if attack is launched immediately after network deployment, the system will not be able to detect that attack [29]. Also the numbers of EM nodes were not specified for specific number of sensor nodes and the proposed method is focused only on static network.

C. Statistical method

In statistical approaches the data associated with certain activities of the nodes in network is studied and recorded by researchers. For example monitor the normal packet transmitted between the nodes or monitor resource depletion of the nodes like CPU usage. Then the adversary or compromised node is detected by comparing the actual behavior with the threshold value which used as reference, if any nodes exceed that value is considered as an intruder.

Chen, et al [3], proposed statistical GRSh (Girshick-RubinShyriaev)–based algorithm for detecting malicious nodes in wireless sensor network. Base station calculates the difference of CPU usage of each node after monitoring the CPU usage of each node in fixed time. Base station would identify whether a node is malicious or not after comparing the difference of CPU usage with the threshold.

Dynamic trust management system was proposed by Roy et al [23] to detect and eliminate multiple attacks such as sinkhole attack. Each node calculates the trust of its neighbor node based on experience of interaction; recommendation and knowledge then sends to base station. The base station decided which node is sinkhole after it received several trust values from other nodes. Therefore the trust value of the node which falls beyond the normal value 0.5 is considered as sinkhole attack [23].

D. Hybrid based intrusion detection

The combination of both anomaly and signature based or misused based is used in this approach. The false positive rate which produced by anomaly based is reduced in this approach due to the use of both method. Also the advantage of this approach is to be able to catch any suspicious nodes which their signature is not included in detection database.

Coppolino and Spagnuolo [6] proposed hybrid Intrusion detection system to detect sinkhole attack and other attacks. They used detection agent which was responsible for identifying sinkhole attack. The hybrid intrusion detection was attached to sensor node and share resource of that node. The suspicious nodes were inserted to the blacklist based on anomalous behavior after analyzed the collected data from neighbors. Then that list is sent to central agent to make final decision based on feature of attack pattern (misused based). Similar to solution proposed by Tumrongwittayapak and

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

Varakulsiripunth [29], it was designed for static wireless sensor network.

E. **Key management**

In key management approach the integrity and authenticity of packet travels within the network is protected by using encryption and decryption key. Any packet transmitted in the network is added with another message in a way that to access that message requires a key and any small modification of the message can be easily detected. Those keys also help nodes to check if the message comes from base station and check the authenticity of the message.

Papadimitriou et al [21] proposed a cryptographic approach in routing protocol to address the problem of sinkhole attack. Each node obtained public key which used to verify if the message comes from base station. They also used pair of public and private keys for authentication and sign data message. All keys were uploaded offline before the network was deployed. Their techniques prevented any node to hide its ID and any packet forgery between nodes in the network. This protocol is focused on resistance to sinkhole attack but not to detect and eliminate it.

Meanwhile, Fessant et al [10] proposed two protocols which used cryptographic method to increase the resilience of sinkhole attack. Both protocols prevent malicious node from lying about their advertised distances to base station. However, they did not show the memory usage of their protocols and message size.

The summary of existing works using the previously described approaches is shown in Table 1.The summary covers evaluation results of proposed solution and their limitations

Table 1: Existing works on Sinkhole detection

Approach	Proposed Solution	Result	Limitations/Advantages
Rule Based. Krontiris et al 2007 [16]	They extended their IDS which can detect sinkhole attack.	• the success of intrusion detection system depend on the increase number of watchdog • When the network density increase the false negative rate decrease.	**Limitations** • Memory and network overhead was created. • They used MintRoute protocol • Node impersonation was the focus of the rules. **Advantages** • More secure and robust measure can be developed based on valuable principle they develop**.**
Rule Based. Krontiris et al 2008 [15]	They proposed detection rules that will keep aware legitimate node the existing of attack.	• They show how vulnerabilities of MultihopLQI can be exploited by sinkhole node and suggest the rules which make the protocol more resilient.	**Limitation** • They did not show practically how those rules can prevent attack. • All the rules are only detecting attack but cannot give ID of sinkhole node. • They assume attacker has the same power as normal node and can capture sensor node and change the internal state.
Anomaly based. Tumrongwittayapak, C and Varakulsiripunth, R 2009 [29]	They proposed detection solution based on received signal strength indicator(RSSI) Their proposed	• For 0 to 40% percentage of message drop the detection rate is 100% • False positive rate was 0 for 0-40% of message drop but increase when percentage drop increase • The same applied to false negative rate with the more	**Limitation** • They assume sensor network are static • No instant attack • Base station remain 0,0 position • Base station and extra monitor node are physically protected. • Their proposed solution can not detect attack if it happened

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

	solution required support from extra monitor node	message drop the more negative rate.	instantly after network deployment.
Anomaly based. Choi et al 2009 [4]	They proposed method that can detect sinkhole attack that used LQI (link quality indicator).	• The probability of detection increase when number of detector nodes increase • detection rate increase when detector node increase • The false positive rate depend on extent of tolerance value (constant value which will show if changes is beyond abnormal)	**Limitations** • All sensor node have no mobility • The detection of sinkhole occurs when detector node is between sinkhole node and source node and sinkhole and base station • The detector nodes have high source of energy than sensor nodes **Advantage** • Detector node communicate themselves through exclusive channel
Anomaly based. Sharmila, S. and Umamaheswari, G. 2011. [24]	-They proposed message digest algorithm to detect sinkhole node.	• The results show the algorithm worked well when malicious nodes are below 50% • False positive rate was 20%(due to packet drop) that figure obtained when malicious node reach 50 • False negative error was 10% but was increasing when malicious node reach above 40	**Limitation** • Network throughput, overhead and communication cost was not calculated • The performance was not good when there is node collision, limited transmitted power and packet drops • Only one advertisement is considered at a time, after computation another take place **Advantage** • The algorithm achieve data integrity and authenticity
Key Management. Papadimitriou et al 2009 [21]	-They proposed two RESIST protocols which increase resilience to sinkhole attack in WSN	-Results show that RESIST-0 has high resilience to sinkhole attack (it does not allow node to lie about their distance to base station) than other protocol	**Limitation** • Resist-0 is very expensive it require two additional message to a packet • In their simulation message losses and collusion were not considered • Collusion node has impact on RESIST-0 not in RESIST-1 • Their routing algorithm relying on tree-based topology construction • Route tree is built by hop distance **Advantage** • RESIST-1 prevent malicious nodes from changing their advertised distance to the sink more than one hop • RESIST-0 completely stops any

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

			lying about distance.
Statistical based Chen et al 2010 [3]	They develop an algorithm which detect sinkhole attack and identified intruder.	• From first simulation the detection time increase when threshold (CPU value) become bigger • Also the false positive rate decrease when threshold become bigger. • From the second simulation the detection time did not change too much but the false positive rate increase due to increase in traffic	**Limitation** • Base station makes the final decision on which node is malicious • No results on the network overhead • The scheme will not detect attack if it launch instantly after deployed • Assumption-base station is trustworthy and it participates in detection system. **Advantages** • Their algorithm showed that it can detect malicious node in short time with low false positive rate
Hybrid base Coppolino et al 2007 [6]	They proposed intrusion detection system which was able to protect critical information from attacks directs from its WSN.	• Detection rate was 95-97% when malicious node modified sensor packet. • Detection rate was 93-96% when malicious node modified the r • False positive rate is 3% • IDS usage in real sensor network was 734bytes (RAM) and 3208bytes (ROM)	**Advantage** • Their solution satisfied the available resource in sensor nodes • Their solution proved to detect sinkhole attack • They used both anomaly and misuse based method
A non cryptographic Sheela, D et al 2011 [25]	They proposed scheme which used mobile agent to defend against this attack	• Probability of detecting sinkhole is decrease when nodes increase • Node average energy decrease as time goes up because of storage information. • The algorithm create high network overheads	**Limitation** • Mobile wireless sensor network • No specification of exactly number of MA(mobile agent) in network • Matrix method is very complex with relate to available resources • MA communicate with sensor nodes at active mode only **Advantage** • MA used dummy data to detect modification • MA has sufficient power to run its activities

V. DISCUSSION

From the Table 1, it shows most approaches managed to detect and prevent sinkhole attack in WSN.

Rule based approaches managed to detect sinkhole attack but it creates memory and network overhead. This approach did not give the ID of sinkhole node after detection of attack. All the rules focus on the node impersonation.

Anomaly based approach also manage to detect sinkhole attack but they just focus on static wireless sensor network. This approach created high false positive rate when there was high message dropping.

Key management was another approach which focused on resistance to sinkhole attack but not to detect and eliminate it.

Statistical based approach managed to detect sinkhole attack but they did not give result of the network

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

overhead. Also this approach cannot detect an instant attack after WSN is deployed. False positive rate were the main drawback to this approach.

Hybrid based intrusion detection approach used the combination of both anomaly and signature-based. This approach detected sinkhole attack but was designed for static WSN. It produced less false positive rate.

A non cryptographic is another approach which detected sinkhole attack but it created high network overhead.

All the approaches managed to detect, identify and provided resistance to sinkhole attack. The major drawbacks produce by those approaches includes high network and memory overhead, create high false positive rate and some were not able to work on mobile WSN.

VI. CONCLUSION AND FUTURE WORK

Based on existing works most researchers are trying to look for ICT solutions for detecting, identifying and providing resistance to sinkhole attack in wireless sensor network. Researchers used intrusion detection scheme based on anomaly-method, other used rule based and key management to detect and identifying the sinkhole nodes. Majority of researches struggled with security challenges corresponding with availability of resources and mobility of wireless sensor nodes. Some provided solution for only static and few on mobile network. Very few researchers managed to validate their security system using real wireless sensor network. Also some of results showed low detection rate, high network overhead and high communication cost. The future solution should focus on reducing high network overhead, computational power, increase detection rate and that system must be validated in real sensor network. Through this kind of validation, it will be easy to check if their solutions meet the available resources of WSN, such as memory capacity.

ACKNOWLEDGEMENT

We like to thank almighty God for giving us strength during the all period of the research. We would to like to express our sincere appreciation to other colleague in our department for their advice.

REFERENCES

[1]. Abhishek Pandey and R.C. Tripathi. (2010). A Survey on Wireless Sensor Networks Security, International Journal of Computer Applications (0975 – 8887) Volume 3 – No.2.

[2]. Changlong Chen, Min Song, and George Hsieh (2010) Intrusion detection of Sinkhole attack in large scale Wireless Sensor Networks, 978-1-4244-5849-3/10/ IEEE.

[3]. Chen, C., Song, M. and Hsieh, G. (2010). Intrusion Detection sinkhole attack in large scale wireless sensor network

[4]. Choi, G. B., Cho, J. E., Kim,H. J., Hong, S. C. and Kim, H. J. (2008).A sinkhole attack detection mechanism for LQI based mesh routing in WSN.

[5] Chun-ming Rong , Skjalg Eggen, Hong-bing Cheng. (2011). A Novel Intrusion Detection Algorithm for Wireless Sensor networks. Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 2nd International Conference.

[6] Coppolino, L., D'Antonio, S., Romano, L., and Spagnuolo, G.(2010). An intrusion detection system for critical information infrastructures using WSN technologies.

[7] Daniel Dallas, Christopher Leckie, Kotagiri Ramamohanarao. (2007). Hop-Count Monitoring: Detecting Sinkhole Attacks in Wireless Sensor Networks, ICON 15th IEEE International Conference.

[8]. David Martins and Hervé Guyennet. (2010) Wireless Sensor Network Attacks and Security Mechanisms: A Short Survey, 13th International Conference on Network-Based Information Systems

[9]. Edith C. H. Ngai, Jiangchuan Liu and Michael R. Lyu. (2006). On the Intruder Detection for Sinkhole Attack in Wireless Sensor Networks, reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC proceedings

[10].Fessant, F., Papadimitriou, A., Viana, A.,Sengul, C. and Polamar, E. (2011) A sinkhole resilient protocol for wireless sensor network: Performance and security analysis.

[11]. G.H. Raghunandan, B.N. Lakshmi. (2011.). A Comparative Analysis of Routing Techniques for Wireless Sensor Networks, Proceedings of the National Conference on Innovations in Emerging Technology Kongu Engineering College, Perundurai Erode, Tamilnadu, India. PP.17-22.

[12]. Jaydip Sen. (2009). A Survey on Wireless Sensor Network Security, International Journal of Communication Networks and Information Security (IJCNIS) Vol. 1, No. 2

[13] Kalpana Sharma and M K Ghose. (2010) Wireless Sensor Networks: An Overview on its Security Threats, IJCA Special Issue on "Mobile Ad-hoc Networks" MANETs

[14]. Krontiris,I., Dimitriou,T., Giannetsos,T. and Mpasoukos, M. (2008). Intrusion Detection Sinkhole Attacks in Wireless Sensor Network. LNCS 4837, pp. 150-161

[15]. Krontiris, I., Giannetsos, T. and Dimitriou, T. (2008). Launch Sinkhole Attack in Wireless Sensor Network; the Intruder Side.

[16] Krontiris, I. Dimitrou, T. Freiling, F.C. (2007). Towards intrusion detection in wireless sensor networks. Proceedings of the 13th European Wireless Conference, Paris, France, April 2007.
[17]. Liping Teng and Yongping Zhang. (2010). SeRA: A Secure Routing Algorithm against Sinkhole Attacks for Mobile Wireless Sensor Networks, IEEE Second International Conference on Computer Modeling and Simulation.
[18]. Ngai, E., Liu, J and Lyu, M. (2007) An efficient intruder detection algorithm against sinkhole attack in wireless sensor network.
[19]. Ngai, E., Liu, J. and Lyu, M. (2006). On intruder detection for Sinkhole attack in Wireless Sensor Network.
[20]. P. Samundiswary, D.Sathian and P. Dananjayan. (2010). Secured greedy perimeter stateless routing for wireless sensor networks, International Journal of Ad hoc, Sensor & Ubiquitous Computing(IJASUC) Vol.1, No.2.
[21]. Papadimitriou, A., Fessant, L. F. and Sengul, C. (2009). Cryptographic protocols to fight sinkhole attacks on tree based routing in WSN.
[22]. Pathan, K., Al-S. (2011).Security of Self-Organizing Networks-MANET, WSN, VANET, WMN. International Standard Book Number-13: 978-1-4398-1920-3 (Ebook-PDF). Tayrol and Francis Group http://www.taylorandfrancis.com
[23] Roy, D.S., Singh, A.S. and Choudhury, S. (2008). Countering Sinkhole and Blackhole Attacks on Sensor Networks using Dynamic Trust Management.
[24]. Sharmila, S. and Umamaheswari, G. (2011). Detection of sinkhole attack in WSN using message digest algorithms.
[25]. Sheela, D., Kumar, N., and C Dr. Mahadevan, G.C.(2011). A non-Cryptographic Method of Sinkhole Attack Detection in Wireless Sensor Networks.
[26]. Suman Deb Roy, Sneha Aman Singh, Subhrabrata Choudhury, and N. C. Debnath. (2008). Countering Sinkhole and Black hole Attacks on Sensor Networks using Dynamic Trust Management", 978-1-4244-27 IEEE
[27]. Teng, L., and Zhang, Y. (2010). Secure Routing Algorithm against Sinkhole attack for Mobile Wireless Sensor Network
[28]. Tumrongwittayapak, C and Varakulsiripunth, R. (2009). Detecting sinkhole attack and Selective forwarding attack in WSN.
[29] Tumrongwittaya and Varakulsiripunth. (2009). Detection of Sinkhole attack in Wireless Sensor Networks, ICROS-SICE International Joint Conference.

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

Decomposition and denoise of an image using singular value decomposition

Viji Paul
Department of Information Technology
Salalah College of Technology, Sultanate of Oman

D. Nagarajan
Department of Information Technology
Salalah College of Technology, Sultanate of Oman

P.Selvaraju
Department of Computer Science & Engineering
Vel Tech Multi Tech Dr.RR and Dr.SR Engineering College, Avadi, Chennai, India

Abstract **- Image decomposition is now essential for transmission and storage in database. Singular Value Decomposition is a decomposition technique for calculating the singular values, pseudo-inverse and rank of a matrix. The conventional way of doing this was to convert a matrix to row echelon form. The rank of a matrix is then given by the number of nonzero rows or columns of the echelon form. Singular value decomposition is one of the methods to compress and denoise the images. The main focus of this paper is to decompose and denoise an image using singular value decomposition.**

Key words :SVD, decomposition, de noise .

I. INTRODUCTION

A digital image is generally encoded as a matrix of grey level or color values. Each pair (i,u(i)),where u(i) is the value at i is called a pixel. The image accuracies are categorized as noise and blur. Noise is the perturbation of the image and blur is intrinsic to image acquisition system. A good quality image has a standard deviation of about 60. Image compression is one of the applications in SVD. Consider some matrix A with rank 1000; that is, the columns of this matrix span a 1000 dimensional space. Encoding this matrix on a computer is going to take quite a lot of memory. We might be interested in approximating this matrix with one of lower rank. An image is a section of random access memory that has been copied to another memory or storage location. Dimensionality reduction is a noise reduction process. Thus, SVD belongs to a class of dimensionality reduction techniques that deal with the uncovering of hidden data structures. If matrix A is in the form of $A = USV^T$, where U is a matrix whose columns are the eigenvectors of the AA^T matrix. These are termed the left eigen vectors. S is a matrix whose diagonal elements are the singular values of A. This is a diagonal matrix, so its nondiagonal elements are zero by definition. V is a matrix whose columns are the eigenvectors of the A^TA matrix. These are termed the right eigenvectors. V^T is the transpose of V [1].

$A^* = U^* S^* V^{T*}$. This process is termed dimensionality reduction, and A* is referred to as the Rank k Approximation of A or the "Reduced SVD" of A. The top k singular values are selected as a mean for developing the representation of A, which is now free from noisy dimensions [2].

II. DATA BASE

Pixel of the image is 675 rows, 900 columns and 3 colours. Signal processing aims at extracting information from the raw signal. The difficulty in reaching this goal depends both on the characteristics of the noise-free signal and the noise. The signal-to-noise-ratio (SNR) is the ratio of the strength of the signal and the strength of the noise. The higher the ratio the easier it is to extract information and the more reliable are the results. SNR is the ratio, of the mean and the standard deviation of the measured signal.

SNR = $\bar{x}$ / s

Calculate the signal as the mean of pixel values. Calculate the Pnoise and the standard deviation or error value of the pixel values. Take the ratio or you may use SNR=10log10(Psignal / Pnoise) to express the result in decibel. 0.8620.The above image with standard deviation 66.04.

III. MODEL DESCRIPTION

The singular value decomposition closely associated to the companion theory of diagonal in a symmetric matrix. If A is a symmetric real n x n matrix there is an orthogonal matrix V and a diagonal D such that,

$$A=VDV^T.$$

Here the columns of V are latent vectors for A and diagonal entries of D are eigen values of A for Singular Value Decomposition begin with m x n real matrix. There

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

are orthogonal matrices U and V and a diagonal matrix S, such that

$$A = USV^T$$

Here U is m x m and V is n x n, so that S is rectangular with the same dimensions as A .The matrix S can be formatted to be non-negative and in order of decreasing order. The columns of U and V are called left and right Singular vectors for A[5],[8].

IV. Analysis

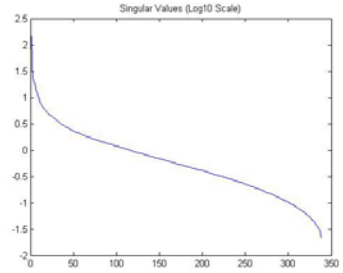
The analysis is done in MATLAB software. The image can be converted to black and white, and then the image is treated as a matrix.

```
a=imread('ncduck.jpg');
imshow(a)
[m,n,k]=size(a)
ncduck = rgb2gray(imread('ncduck.jpg'));
ncduck = im2double(imresize(ncduck, 0.5));
[U, S, V] = svd(ncduck);
```

From the above algorithm the image is converted to the singular value decomposition.

```
sigmas = diag(S);
figure; plot(log10(sigmas)); title('Singular Values (log10 Scale)');
```

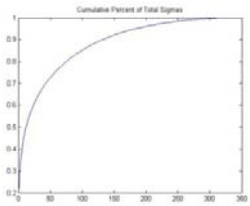
The above algorithm shows the singular value of the image, which is a base 10 log scale.



Approximately first thirty ranks get the largest singular values.

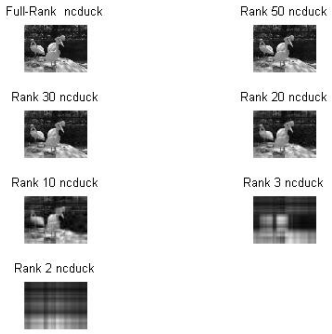
```
figure; plot(cumsum(sigmas) / sum(sigmas));
title('Cumulative Percent of Total Sigmas');
```

The above algorithm lines shows the cumulative percentage of the singular value .



```
figure; subplot(4, 2, 1), imshow(ncduck), title('Full-Rank ncduck');ranks = [50, 30, 20, 10, 3,2];
for j = 1:length(ranks)
approx_sigmas = sigmas; approx_sigmas(ranks(j):end) = 0;
ns = length(sigmas);
approx_S = S; approx_S(1:ns, 1:ns) = diag(approx_sigmas);
approx_ncduck = U * approx_S * V';
subplot(4, 2, j + 1), imshow(approx_ncduck),
title(sprintf('Rank %d ncduck', ranks(j)));
end
```

Approximation rank of the image as follows it shown below:



From the above images it shows that, higher the singular value, better the quality of the image. Singular value decomposition compressed a 675x900 pixel image into a 675 x 675 for U, 30 x 30 singular value and a 30 x 900 matrix. Singular values can be used to highlight which dimensions are affected the most when a vector is multiplied by a matrix.

V. De noise.

The singular value decomposition ,

$A = USV^T$, $A^T = (USV^T)^T$ The discrete noisy image $v = \{v(i) / i \in I\}$, $\omega(p,q)$ depends on the similarity between the pixels p and q and satisfies the condition of the probability $0 \le \omega(p,q) \le 1 \; and \; \sum_q \omega(p,q) = 1$

[4]Computing the similarity between the images pixels will depend on the similarity of the intensity grey level of vector p referred as black and q as white. Then, $p = p^T U_K S_K^{-1}$ and $q = q^T U_K S_K^{-1}$. The similarity of p and q $sim(p,q) = sim(p^T U_K S_K^{-1}, q^T U_K S_K^{-1})$ .The large weight of similarity windows are similar and smaller because of the intensity grey values in the similarity windows are varying. White noise alters the distance between windows in a uniform way. Impixel region creates a Pixel Region associated with the image displayed in the current figure, called the target image. The Pixel Region tool opens a separate figure window containing an extreme close-up view of a small region of pixels in the target image [7].

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

VI. COMPRESSION MODEL

Use the compress button to bring up the Wavelet Packet 2-D Compression window. Select the remove near zero option from the Select thresholding method. Threshold of the image is 4.995. [3],[6].

Notice that the default threshold (1.5 higher degree) provides about 43.61% compression while retaining virtually all the energy of the original image. Depending on criteria, it may be worthwhile experimenting with more aggressive thresholds to achieve a of compression. This can be considered a precompression step in a broader compression system. Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) are used to compare the squared error between the original image and the reconstructed image. There is an inverse relationship between PSNR and MSE. Higher PSNR value indicates the higher quality of the image.

PSNR of the compressed image is 28.02, which means, 28% of the noise removed from the image.

VII. CONCLUSION

The results show that, one of the main decomposition approach based on singular value decomposition for adaptive noise is de-noising. Experimental results are proposed for performance of PSNR on visual effect in color images, even in presence of high ratio of noise. High PSNR value reveals good image quality and less error introduced in to the image. In case of loss less compression PSNR will be high. Further the work can be extended to 3D framework image denoesing and image restoration.

REFERENCES

[1] Chan.T , “An improved algorithm for computing the singular value Decomposition”, ACM Transactions on Mathematical Software, Vol 8, No. 1 (1982).
[2] Berry, M., S. Dumais, and G. O'Brien .,”Using Linear Algebra for Intelligent information” Retrivel. SIAM Review 37(4): 573-595. (1995).
[3] T. F. Chan and H. M. Zhou, “Total variation improved wavelet thresholding in image compression”, in Proceedings of the IEEE International Conference on Image Processing, Vol. 2,Vancouver, BC, Canada, 2000, pp. 391–394.
[4] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger,Universal Discrete denoising: Known channel, IEEE Trans. Inform. Theory, 51 (2005), pp. 5–28.
[5] E. Tadmor, S. Nezzar, and L. Vese, A multiscale image representation using hierarchical (BV,L2) decompositions, Multiscale Model. Simul., 2 (2004), pp. 554–579.
[6] H. Choi and R. G. Baraniuk, (2004) “Multiple wavelet basis image denoising using Besov ball projections,” IEEE Signal Processing Letters, Vol. 11, No. 9, 2004, pp. 717–720.
[7] Z. Tongzhou; W. Yanli R. Ying; L. Yalan “Approach of Image Denoising Based on Discrete Multi-wavelet Transform International Workshop on Intelligent Systems and Applications, 2009, pp: 1-4.
[8] D.Nagarajan ,P.Sunitha, V.Nagarajan and V.Seethalekshmi (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010.

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

An improvement on fragmentation in Distribution Database Design Based on Knowledge-Oriented Clustering Techniques

VAN NGHIA LUONG
Department of Information Technology
Pham Van Dong University
Quang Ngai, Viet Nam

HA HUY CUONG NGUYEN
Department of Information Technology
Quang Nam University
Quang Nam, Viet Nam

VAN SON LE
Da Nang University of Education
Da Nang University
Da Nang, Viet Nam

***Abstract*—The problem of optimizing distributed database includes: fragmentation and positioning data. Several different approaches and algorithms have been proposed to solve this problem. In this paper, we propose an algorithm that builds the initial equivalence relation based on the distance threshold. This threshold is also based on knowledge- oriented clustering techniques for both of horizontal and vertical fragmentation. Similarity measures used in the algorithms are the measures developed from the classical measures. Experimental results carrying on the small data set match fragmented results based on the classical algorithm. Execution time and data fragmentation significantly reduced while the complexity of our algorithm in the general case is stable.**

Keywords— Vertical Fragmentation; Horizontal Fragmentation; Similarity Measure; Clustering Techniques knowledge-oriented clustering techniques.

I. INTRODUCTION

In distributed computing environments, each unit of data (item) which is accessed at the station, (site) is not usually a relationship but part of the relationship. Therefore, to optimize the performance of the query, the relations of global schema are fragmented into items.

There are several types of data fragmentation that are use vertical fragmentation, horizontal fragmentation, mixed fragmentation and derived fragments. Two classical algorithms associated with horizontal fragmentation and vertical fragmentation are PHORIZONTAL and BEA respectively [11].Many authors have proposed solutions improved the above two algorithms, as Navathe, et al., (1984) [13], Chakravarthy, et al., (1994) [3]..., However, the complexity of this algorithm is quite large, with vertical fragmentation problem is $O(n^2)$, where n is the number of attributes and horizontal fragmentation is $O(2^m)$, where m is the number of records [9],[11].

In recent years, several authors have incorporated to solve the problem of fragmentation and positioning, by using the optimal algorithms [5-6], [10] or using the heuristic method [4], [7]. The execution time of these algorithms is remarkably smaller than the classical algorithm.

The used technical association rules in data mining to vertical fragmentation has been mentioned in [8]. However, the data mining techniques do not attract many authors.

In this paper, we use knowledge-oriented clustering techniques for vertical and horizontal fragmentation problem. The measure of similarity was developed based on from the available measure of the classical algorithms in data mining.

In the clustering algorithm based on knowledge-oriented, we propose an algorithm that builds the initial equivalent relation based on the distance threshold. This approach differs from the previous works proposed by Hirano et al.[10], [14] and Bean et al.,[2], [14-16] in that the proposed algorithm automatically determines the number of clusters based on the data set of survey.

The paper is organized as follows: section 2 presents a brief overview of the basic concepts. We detail with the proposed vertical and horizontal fragmentation algorithms, in section 3 and section 4, respectively. We then discuss the main contributions of proposed approach in section 5.

II. BASIC CONCEPTS

A. *Vertical fragmentation*

Vertical fragmentation is the collective decay properties of the relational schema R into the sub schema R_1, R_2, ..,R_m, such that each attribute in these sub schemas is often accessed together.

To show how often the same queries together, Hoffer and Severance introduced the concept attribute affinity [11].

If $Q = \{q_1, q_2, .., q_m\}$ is a set of applications, $R(A_1, A_2, .., A_n)$ is a relational schemas. The relationship between q_i and attributes A_j is determined by using the values:

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

$$use(q_i, A_j)=\begin{cases}1, & A_j \text{ is engaged in } q_i\\ 0, & A_j \text{ is not engaged in } q_i\end{cases} \quad (1)$$

Put $(A_i, A_j) = \{q \in Q \mid use(q, A_i).\ use(q, A_j) = 1\}$. Attribute affinity between A_i and A_j is:

$$Aff(A_i, A_j)= \sum_{q \in Q(A_i, A_j)} \left(\sum_{\forall S_l} ref_l(q) * acc_l(q) \right) \quad (2)$$

In particular, refl(q): the number of pairs of attributes (A_i,A_j) is referenced in the application q at station S_l; accl(q): frequency of access to applications q in station S_l. BEA algorithm consists of two main phases:

1) Permutations row, column affinity matrix of attribute to obtain the cluster affinity matrix (CA) which has global affinity measure AM (global affinity measure) [11] is the largest.

2) Find the partition of the set of attributes from the matrix CA by exhaustive method, so that:

Z= CTQ *CBQ – COQ2 is the maxima, with:

$$CTQ= \sum_{q \in TQ} \sum_{\forall Sj} ref_j(q_j) acc_j(q_i)$$

$$CTQ= \sum_{q \in TQ} \sum_{\forall Sj} ref_j(q_j) acc_j(q_i)$$

$$COQ= \sum_{q \in OQ} \sum_{\forall Sj} ref_j(q_j) acc_j(q_i)$$

TABLE I. CLUSTER AFFINITY MATRIX CA

	A_1	A_2	..	A_i	A_{i+1}	..	A_n
A_1							
..			*TA*				
A_i							
A_{i+1}							
						BA	
A_n							

In which,

$AQ(q_i)= \{A_j \mid use(q_i, A_i)=1\}$;

$TQ=\{q_i \mid AQ(q_i) \subseteq TA\}$;

$BQ= \{q_i \mid AQ(q_i) \subseteq BA\}$;

$OQ=Q \setminus \{TQ \cup BQ\}$

The complexity of the algorithm is proportional to n^2.

B. Horizontal Fragmentation

Horizontal fragmentation divides set records into a smaller set of records. Horizontal fragmentation is based on the query conditions, which are expressed through simple predicates of the form: A_j θ<value>.

Set Pr = $\{Pr_1, Pr_2, .., Pr_k\}$ is a set of simple predicates extracted from a set of applications. A conjunction of the predicates, which is built from P_r will have the form:

$$P_1^* \wedge P_2^* \wedge ... \wedge P_n^* \quad (3)$$

Where P_i^* is a predicate, which received one of P_i or $\neg P_i$ values.

PHORIZONTAL algorithm uses the conjunction of the predicates $P_1^* \wedge P_2^* \wedge ... \wedge P_n^*$ to find the conditions for horizontal fragmentation of data [9]. The relation r(R) will be fragmented into $\{r_1(R), r_2(R),..,r_k(R)\}$, with $r_i(R) = \sigma F_i(r(R))$, $1 \leqslant i \leqslant k$; F_i is a predicate, which forms the conjunction of the primary predicates [9].

C. Information systems and the inability to distinguish relationship

- The information system is a pair of SI = (U, A), where U is a finite set of objects $U=\{t_1, t_2, .., t_n\}$, A is non-empty finite set of attributes.
- An equivalent relation (A binary relations satisfy properties reflective, symmetric and transitive) defined on U is called an inability to distinguish relationship (irrespectively relationship) on U.

D. Clustering algorithm Knowledge-Oriented

Clustering algorithm Knowledge-Oriented based on rough set theory was first proposed by the authors Shoji Hirano, et al., [10], [15]. This is a clustering algorithm automatically determines the number of clusters based on the survey data [12]. The main idea of this clustering algorithm consists of 2 phases:

1) Created of equivalence relation on the set of object clustering.

2) Editing of the equivalence relation using a threshold Tk based on the measure irrespective. This iterative process will update Tk a best clustering results is obtained Using this algorithm to data fragmentation, we have proposed initial equivalence relation based on the average distance between objects.

Clustering algorithm Based on knowledge orientation, so we propose as follows:

Input: U= the set of objects to be clustered.

(Each object must be describe the information needed to construct a similar measure).

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

Output: The clusters (corresponding to the fragment of data).

Method:

Step 1: Construct a matrix of similarities $S=S(t_i, t_j)$ between all pairs of objects (t_i, t_j);

Step 2: Specify a initial ability to distinguish relationship R_i for each object. Synthesis to get an initial clustering;

Step 3: Construct ability to distinguish matrix $\Gamma=\gamma(t_i, t_j)$ to assess the quality of clustering;

Step 4: Modify the clusters by the inability to distinguish relationship R_j^{mod} for each object to achieve the revised clustering;

Step 5: Repeat steps 3 and 4 until a stable clustering is obtained.

The inability to distinguish relationship corresponds to the i^{th} attribute:

$R_i = \{(t_i, t_j \in U \text{ x } U : d(t_i, t_j) \leqslant Tk_j, \text{ with } j = 1,2,\ldots,n\}$.

Where $d(t_i, t_j)$ is the distance between two participants clustering.

The threshold Tk_j is determined as follows:

$$Tk_i \left[\sum_{j=1,\ j \neq i}^{n} (1 - s(t_i, t_j)) \right] /(n-1) \quad (4)$$

With $s(t_i,t_j)$ is the similarity measure of two objects t_i, t_j.

III. VERTICAL FRAGMENTATION PROBLEM BASED ON KNOWLEDGE- ORIENTED CLUSTERING TECHNIQUES

Vertical fragmentation problem is converted to the clustering problem, based on the following concepts:

A. Attribute and the reference feature vector

Definition 1: The reference measure of transaction q_i on attribute A_j, denoted by $M(q_i, A_j)$:

$$M_{ij} = M(q_i, A_j) = use(q_i, A_j) * f_i$$

In which M_{ij} is the frequency with which transactions q_i reference to attribute A_j. With f_i is the frequency of transactions q_i and $use(q_i, A_j)$ is defined by formula (1).

Definition 2: VA_j reference feature vector of attribute A_j with reference transactions $(q_1, q_2, ..,q_m)$ is defined as follows:

	q_1	q_2	…	q_m
$VA_j =$	M_{1j}	M_{2j}	…	M_{mj}

B. The similarity measure of two properties

Definition 3: The similarity measure of two attributes A_k, A_1 has two feature vectors corresponding to the reference transactions $(q_1, q_2, ..,q_m)$:

$VA_k = (M_{1k}, M_{2k}, ..,M_{mk})$

$VA_1 = (M_{1l}, M_{2l}, ..,M_{ml})$

Is determined by the cosine measure:

$$s(A_k, A_l) \frac{VA_k * VA_l}{\|VA_k\| * \|VA_l\|} = \frac{\sum_{i=1}^{m} M_{ik} * M_{il}}{\sqrt{\sum_{i=1}^{m} M_{ik}^2} * \sqrt{\sum_{i=1}^{m} M_{il}^2}} \quad (5)$$

C. Vertical fragmentation based on knowledge- oriented clustering technique

To illustrate the vertical fragmentation algorithm based on knowledge-oriented clustering techniques. We use the assumption of examples about vertical fragmentation problem based on BEA algorithm is presented in [1], [8]:

The set of attributes $A_t = \{A_1, A_2, A_3, A_4\}$

The set of transactions $Q = \{q_1, q_2, q_3, q_4\}$. The matrix used:

	A_1	A_2	A_3	A_4
q_1	1	0	1	0
q_2	0	1	1	0
q_3	0	1	0	1
q_4	0	0	1	1

The frequency of application execution with a set of transactions $\{q_1, q_2, q_3, q_4\}$, and $F = \{f_1, f_2, f_3, f_4\} = \{45, 5, 75, 3\}$.

From the assumption, we have the reference feature vectors:

	q_1	q_2	q_3	q_4
$VA_1 =$	45	0	0	0
$VA_2 =$	0	5	75	0
$VA_3 =$	45	5	0	3
$VA_4 =$	0	0	75	3

The similar matrix $S_{4x4} = (s(A_k, A_l))$ k=1,4; l=1,4

	A_1	A_2	A_3	A_4
A_1	1	0	0.9918	0
A_2		1	0.0073	0.9970
A_3			1	0.0026
A_4				1

The result of the vertical fragmentation algorithm based on the clustering algorithm towards knowledge-oriented.

Cluster	Set of attributes
1	$\{A_1, A_3\}$
2	$\{A_2, A_4\}$

This fragmentation results correlate with the results of the vertical fragmentation by algorithm BEA.

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

IV. HORIZONTAL FRAGMENTATION PROBLEM BASED ON KNOWLEDGE-ORIENTED CLUSTERING TECHNIQUE

Similar to vertical fragmentation, assuming conversion of horizontal fragmentation problem from PHORIZONTAL algorithm is based on the concept of the following establishments:

A. Vectorization records of a relationship

Considering the relations r(R) = $\{T_1, T_2, .., T_l\}$, the set of simple predicates extracted from applications on r(R) is Pr = $\{Pr_1, Pr_2, .., Pr_m\}$. Vector binary of records by the rule:

TABLE II. *VECTORIZATION BINARY*

	P_{r1}	P_{r2}	..	P_{rj}	..	P_{rm}
T_1	a_{11}	a_{12}	..	A_{1j}	..	a_{1m}
..		..				..
Ti	a_{i1}	a_{i2}	..	a_{ij}	..	a_{im}
..		..				..
T_l	a_{l1}	a_{l2}	..	a_{lj}	..	a_{lm}

$$\forall a_{ij} = \begin{cases} 1, \text{ if } Ti\ [Pr_j] = true \\ 0, \text{ if } Ti\ [Pr_j] = false \end{cases}$$

B. The similarity measure of two binary vector

Consider two vectors x_i and x_j, that are represented by binary variables. Assuming binary variables have the same weight. We have event tables as Table 3. Where q is the number of binary variables equal to 1 for the two vectors xi and x_j, s is the number of binary variables equal to 0 fo simpler xi but equal to 1 for x_j, r is the number of binary variables equal to 1 for xi but is 0 for x_j, t is the number of binary variables equal to 0 for all vectors x_i and x_j.

TABLE III. EVENT TABLE FOR BINARY VARIABLES

	Object j			
Object i		***1***	***0***	*Sum*
	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

- The difference of two vectors x_i and x_j based on the symmetric binary dissimilarity are:

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t} \quad (6)$$

- The similarity measure between two vectors x_i and x_j is defined by the Jaccard coefficient:

$$sim(x_i, x_j) = 1 - d(x_i, x_j) \quad (7)$$

C. Horizontal fragmentation based on knowledge-oriented clustering techniques

Considering the relations EMP, [11]:

TABLE IV. SAMPLE DATA (EMP) FOR HORIZONTAL FRAGMENTATION

	ENo	EName	Title
T_1	E_1	Jjoe	Elect-Eng
T_2	E_2	M.Smith	Syst-Analyst
T_3	E_3	A.Lee	Mech-Eng
T_4	E_4	J.Smith	Programmer
T_5	E_5	B.Casey	Syst-Analyst
T_6	E_6	L.Chu	Elect-Eng
T_7	E_7	R.David	Mech-Eng
T_8	E_8	J.Jone	Syst-Analyst

Consider two simple predicates:

p_1=(Title > ”Programmer”); p_2=(Title < ” Programmer”), with string comparison rules in alphabetical order. Vectorization records by two predicates p_1 and p_2 are:

TABLE V. VECTORIZATION RECORDS

	p_1	p_2
T_1	1	0
T_2	0	1
T_3	1	0
T_4	0	0
T_5	0	1
T_6	1	0
T_7	1	0
T_8	0	1

D. The result horizontal fragmentation of the relation (EMP)

The result horizontal fragmentation of the relation EMP as Table III based on the clustering algorithm towards knowledge-oriented. We have used the similarity measure defined by formula (7), where $d(x_i, x_j)$ is calculated by the formula (6).

With k=2, we have:

Cluster	The set of records
1	T_1, T_3, T_6, T_7
2	T_2, T_4, T_5, T_8

With k=3, we have:

Cluster	The set of records
1	T_1, T_3, T_6, T_7
2	T_2, T_5, T_8
3	T_4

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

And with k=4, we have:

Cluster	The set of records
1	T_1, T_3
2	T_2, T_5, T_8
3	T_4
4	T_6, T_7

This fragmentation results coincide with the results of the horizontal fragmentation by algorithm PHORIZONTAL.

V. CONCLUSION

In this paper, we used knowledge-oriented clustering techniques for fragmentation problem in distributed database systems. With this solution, we also proposed transforming hypothetical of this problem to hypothetical of clustering problems. Experimental results on the data in [11] that results correlate with the results obtained from the classical fragmentation algorithms PHORIZONTAL and BEA. In addition to experimental data as presented, we also tested on a number of different data set. The results are also similar to the two classical algorithms above. In the future work, we will carry out the Analysis of large data sets that to compare test the usability of the proposed solution.

VI. REFERENCES

[1] Adrian Runceanu (2007), Towards Vertical Fragmention in Distributed Databases, International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering, CISSE 2007.

[2] C.L Bean, C.Kambhampati (2008), Automonous Clustering Using Rough Set Theory, Inter-national Journal of Automation and Computing, Vol.5 (No.1). pp. 90-102. ISSN 1476-8186.

[3] S. Chakravarthy, J. Muthura, R. Varadarajan, and S. Navathe (1994), Anobjective function for vertically partitioning relations in distributed databases and its analysis, Tech Rep. UF-CIS-TR-92-045.

[4] Hui Ma, Klaus-Dieter Schewe and Markus Kirchberg (2006), A Heuristic Approach to Vertical Fragmentation, Proceedings of the conference on Databases and Information Systems IV: Selected Papers from the Seventh International Baltic Conference DB&IS'2006.

[5] I. Lungu, T. Vatuiu, A. G. Fodor (2006), Fragmentation solutions used in the projection of Distributed Database System, Proceedings of the 6th International Conference "ELEKTRO 2006", Edis-Zilina University Publishers.

[6] Jiawei Han, Micheline Kambel (2012), Data Mining: Concepts and Techniques, 3rd ed Mor-gan Kaufmann Publishers.

[7] Eng. Marwa F. F. Areed, Prof.Dr.Ali I.El-Dosouky, Hesham A. Ali (2008), A Heuristic Approach for Horizontal Fragmentation and Allocation in DOODB, INFOS2008, March 27-29, 2008 Cairo-Egypt © 2008 Faculty of Computers & Information-Cairo University.

[8] Narasimhaiah Gorla, Pang Wing Yan Betty (2010), Vertical Fragmentation in Databases Using Data-Mining Technique, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

[9] Shahidul Islam Khan, A. S. M. Latiful Hoque (2010), A New Technique for Database Fragmentation in Distributed Systems, International Journal of Computer Applications (0975 – 8887), Volume 5– No.9.

[10] Shoji Hirano and Shusaku Tsumoto(2001), A Knowledge-Oriented Clustering Technique Based on Rough Sets, Computer Software and Applications Conference, COMPSAC'01.

[11] Tamer O., ValduriezP (1999), Principles of Distributed Database Systems. Prentice Hall Englewood Cliffs, Second Edition, New Jersey 07362.

[12] Yin-Fu Huang, Jyh-her Chen (2001), Fragment Allocation in Distributed Database Design, Journal of Information Science and Engineering, pp. 491-506.

[13] Navathe S, Ceri S, Wiederhold G, Dou J (1984), Vertical partitioning algorithms for database design, ACM Trans Database Syst 9(4).

[14] Hoffer, J.A., And Severance, D.G (1975), The use of cluster analysis in physical database design. In Proceedings 1st International Conference on Very Large Databases (Framingham, Mass).

[15] A. S. Darabant (2005), A New Approach In Fragmentation Of Distributed Object Oriented Databases Using Clustering Techniques, Studia Univ. Babes»{Bolyai, Informatica, Volume L, Number 2.

[16] A.S. Darabant, A. Campan & others (2004), AI Clustering Techniques: A New Approach in Horizontal Fragmenta-tion of Classes with Complex Attributes and Methods in Object Oriented Databases, Proc. of the Int. Conf. on The-ory and Applications of Mathematics and Informatics.

Personal information:

- Full name: **Luong Van Nghia**, year of birth 1964.
- Address: 99 Hung Vuong street, Quang Ngai city, Quang Ngai province, Vietnam.
- Academic degree: Master of Computer Science
- Name of Agency: Faculty of Informatics, Pham Van Dong Universisty
- Phone:: (+84) 913 498 804,
- Graduation year: Master in Computer Science, Hue University of Science
- Research Areas: Data mining, Distributed Database, Embedded System.

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

Visualizing Object-oriented Software for Understanding and Documentation

Ra'Fat AL-msie'deen
Department of Information Technology, Mutah University, Al-Karak, Jordan

***Abstract*—Understanding or comprehending source code is one of the core activities of software engineering. Understanding object-oriented source code is essential and required when a programmer maintains, migrates, reuses, documents or enhances source code. The source code that is not comprehended cannot be changed. The comprehension of object-oriented source code is a difficult problem solving process. In order to document object-oriented software system there are needs to understand its source code. To do so, it is necessary to mine source code dependencies in addition to quantitative information in source code such as the number of classes. This paper proposes an automatic approach, which aims to document object-oriented software by visualizing its source code. The design of the object-oriented source code and its main characteristics are represented in the visualization. Package content, class information, relationships between classes, dependencies between methods and software metrics is displayed. The extracted views are very helpful to understand and document the object-oriented software. The novelty of this approach is the exploiting of code dependencies and quantitative information in source code to document object-oriented software efficiently by means of a set of graphs. To validate the approach, it has been applied to several case studies. The results of this evaluation showed that most of the object-oriented software systems have been documented correctly.**

Keywords- Vsound; software engineering; software documentation; software visualization; software understanding; software maintenance; software evolution; software reuse; change impact analysis; object-oriented source code; reverse engineering.

I. INTRODUCTION

Different studies about the software understanding indicate that programmers rely on good software documentation [22]. Software rapidly becomes very complex when its size increases, which make its development very hard task. The very huge amount of information represented in the software source code, at all granularity levels (*i.e.*, package, class, attribute and method) make understanding and documenting software a very difficult, lengthy, and error-prone task [3]. Moreover, manually-written documentation is not feasible for being incomplete, either because it is very time-consuming to create, or because it must frequently be updated [12]. This paper proposes a new approach called Vsound[1] to understand and document the object-oriented (OO) software by visualizing its source code as a set of graphs at all granularity levels. In order to give a precise definition of the software documentation, Vsound considers that the software documentation is the process of taking software source code and understanding it by visualizing its source code as a set of graphs.

Software visualization is the use of the crafts of typography, graphic design, animation, and cinematography with modern human-computer interaction and computer graphics technology to facilitate both the human understanding and effective use of computer software [13]. Software visualization (*resp.* software documentation) can tackle three different types of aspects of software (*i.e.*, static, dynamic and evolution) [2]. The visualization of the static aspects of software focuses on visualizing software as it is coded. While, the visualization of the dynamic aspects of software represents information about a specific run of the software and helps comprehend program behavior and, at last, the visualization of the evolution of the static aspects of software adds the time factor to the visualization of the static aspect of software. This paper tackles only the visualization (*resp.* documentation) of the static aspects of software.

Software comprehension[2] is the process whereby a software practitioner understands a software artifact using both knowledge of the domain and/or semantic and syntactic knowledge, to build a mental model of its relation to the situation [14]. Software understanding is one of the main software engineering activities. Software understanding is the process of taking software source code and understanding it. Software comprehension is necessary when a programmer migrates, reuses, maintains, documents or enhances software systems. Software that is not comprehended cannot be changed [1]. The domains of software documentation and visualization are driven by the need for program comprehension. Software visualization (*resp.* documentation) is a successful software comprehension way. Software comprehension is an important part of software evolution and software maintenance.

The software maintenance process is the most expensive part of software development. Most of time spent in software maintenance is used to comprehend the software code and the instructions that have to be changed [17]. Software maintenance is the modification of a software product after delivery to correct faults, to improve performance or other attributes, or to adapt the product to a changed environment [16]. The software undergoes modification to source code and related documentation due to a problem or the need for

[1] Vsound stands for Visualizing object-oriented SOftware for UNderstanding and Documentation.

[2] *a.k.a.,* "program understanding" or "source code comprehension".

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

enhancement. The goal is to modify the existing software while preserving its integrity [15].

Software must frequently evolve to adjust to new features (*resp.* environments) or to meet specific requirements. Software system wants to evolve in order to be used longer time. The company often evolves frequent release of new versions of the original software. Each release results in the increase of software system size and complexity. Thus, software implementation that facilitates modify is key for reducing maintenance costs and effort. Software evolution reflects the process of progressive change in the attributes of the evolving entity or that of one or more of its constituent elements [20]. In other words, software evolution is linked to how software systems evolve over time.

Software reuse is important to reduce the cost and time of software development. In order to reuse existing source code there is a need to understand and document the software code. With increasing of software complexity (*i.e.*, increasing count of lines of code) the need for reuse grows. Software reuse is the process of creating software systems from existing software rather than building software systems from scratch [18]. Software reuse helps reduce the development and maintenance effort. In addition, software reuse improves software quality and decrease time-to-market [38].

Software comprehending is very necessary for software changes in the maintenance stage. The changes to the software's code may be affected to another part of the code. These situations make the developer spent more time and effort to find the affected lines of the whole code. Change impact analysis is the process of identifying the potential consequences of a change, or estimate what needs to be modified to accomplish a change [21]. Change impact analysis support program understanding by finding the potential effect or dependency information in source code.

This paper proposes a new approach, which aims to document software systems by visualizing their code. The documentation process is very useful for software understanding, maintenance, evolution, reuse and changes. Documentation process involves the creation of alternative representations of the software, usually at a higher level of abstraction. It also involves analyzing the software in order to determine its elements and the relations between those elements. Software visualization is commonly used in the fields of reverse engineering and maintenance, where huge amount of code need to be understood [23]. Reverse engineering is the process of analyzing a subject system to identify the system's components and their interrelationships and create representations of the system in another form or at a higher level of abstraction [16].

To assist a human expert to document software system, this paper proposes an automatic approach, which generates a set of graphs (*i.e.*, documents) using source code elements of a software system. Compared with existing work that documents source code (*cf.* section *related work*), the novelty of Vsound approach is that it exploits source code dependencies and its quantitative information, to document OO software in an efficient way by means of a set of graphs. Vsound accepts as input the source code of software as a first step. Then, based on the static code analysis, Vsound generates an XML file contains the main source code elements (*e.g.*, package, class attribute and method) and the dependencies (*e.g.*, inheritance) between those elements. Next, Vsound documents the software by extracting a set of graphs based on the source code; each graph considers as a document. The mined documents cover all granularity levels (*i.e.*, package, class, attribute and method) of the source code.

The Vsound approach is detailed in the remainder of this paper as follows. Section 2 briefly presents the background needed to understand the proposal. Section 3 shows an overview of Vsound approach. Section 4 presents the software documentation process step by step. Section 5 describes the experiments that were conducted to validate Vsound proposal. Section 6 discusses the related work, while section 7 concludes and provides perspectives for this work.

II. BACKGROUND

This section provides a glimpse on software documentation and visualization. It also shortly describes dependencies between source code elements which consider relevant to Vsound approach.

The software documentation process aims to generate documents with abstract information based on software source code. The extracted documents are very useful, especially when software documents are missing. The good software documentation process helps the programmer working on software to understand its features and functions. The software documentation process is specific, where it provides all the information important to the person who works on the software at different levels of abstraction. The documentation process aims to translate source code of the software system into a set of documents (*i.e.*, graphs). In reality, several software systems have little to no software documentation, especially the legacy software. Many companies are facing some problems with legacy systems such as: software understanding and software maintenance. The reason behind these problems is the absence of software documentation. Software systems that are not documented hard to be changed [41]. Common examples of such documentation include requirement and specification documents [22]. Vsound provides as output a set of documents describe the source code and its dependencies.

Software visualization is the graphical show of information about the software source code. Software visualization is not a simple process since the amount of information to be included in the graph is may be far bigger than can be displayed. The software visualization tool presents information about the software source code at different levels of abstraction. The visualization tool focuses on displaying different aspects of the source code. It should provide a way to choose and display just particular information based on the software source code. Usually, the visualization process must find out the level of abstraction of the information it presents about the software source code. The software visualization tool must convert the software source code into a graph. It also must able to visualize a huge amount of information regarding the software source

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

code. Moreover, the visualization tool must provide an easy way for navigation [24].

Vsound relies on the source code of software systems to extract a set of documents describing the software. The software source code is the most important resource of information when the software documentation is missing. In order to document existing OO software based on its source code there is a need to extract the main source code elements and dependencies between those elements. Vsound considers that the software implementation consists of OO source code elements and the code dependencies. The main elements of OO source code include: package, class, attribute and method. Vsound also focuses on the method body. The method body consists of a method signature, which represents the access level, returned data type and parameter list (*i.e.,* name, data type and order). It also consists of the body of the method (*i.e.,* local variable, method invocation and attribute access). Vsound focuses on the main dependencies between source code elements such as: inheritance, attribute access and method invocation in the documentation process. Inheritance relation occurs when a general class (*i.e.,* superclass) is connected to its specialized classes (*i.e.,* subclasses). Method invocation relation occurs when methods of one class use methods of another class. While, attribute access relation occurs when methods of one class use attributes of another class.

III. Approach Overview

This section provides the main concepts and hypotheses used in the Vsound approach. It also gives an overview of the software documentation process. Then, it presents the OO source code model. Finally, it shortly describes the example that illustrates the remaining of the paper.

A. Key Ideas

The general objective of Vsound approach is to document the source code of a single software system based on the static analysis of its source code. Mining the main entities of the source code in addition to the source code dependencies is a first necessary step towards this objective. Vsound considers software systems in which software functionalities are implemented at the programming language level (*i.e.,* source code). Vsound also restricts to OO software system. Thus, software functionalities are implemented using OO source code elements such as packages, classes, attributes, methods or method body elements (*i.e.,* local variable, attribute access, method invocation). There are several ways to document the software source code such as generate a descriptive comments that summarize all software classes or methods [12]. Vsound aims to document the software source code as a set of documents (*i.e.,* graphs). The software documentation process via single graph is difficult. Thus, there is a need to document the software source code through several documents with details.

The documentation process must cover all granularity levels of the software source code (*i.e.,* package, class, attribute and method). Vsound focuses on the extracting of three types of documents (*i.e.,* graphs). The first document contains general information about the software source code. This document called the package document. It represents all software packages in addition to the number of classes per package. It also provides quantitative information about a software system. The second document contains information about software classes. This document called the class document. It represents all information about the class, such as the number of attributes and methods in addition to class dependencies. The third document contains information about software methods. This document called the method document. It represents all information about the method, such as the number of parameters and local variables in addition to method dependencies like attribute access and method invocation.

B. The OO Software Documentation Process

Vsound goal is to document the OO software by using the source code of this software. The software documentation process takes the software's source code as its inputs and generates a set of documents as its outputs.

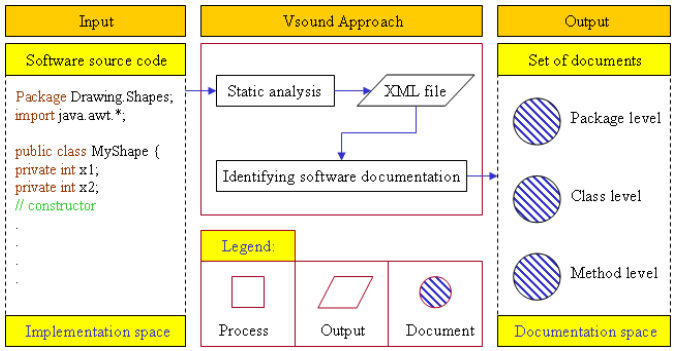


Figure 1. The OO software documentation process.

Vsound approach exploits the main source code elements and the dependencies between those elements in order to document and understand existing software system. Figure 1 shows the software documentation process. The first step of this process aims at extracting the main OO source code elements (*i.e.,* package, class, attribute, method) and their relationships (*i.e.,* inheritance, method invocation and attribute access) based on the static analysis of source code. In the second step, Vsound approach relies on the mined source code to document software at package level. In the third step, Vsound approach documents the software at class level based on the extracted source code. The last step of this process aims at documenting the software at method level. Finally, these documents (*i.e.,* graphs) are used to understand and document the software system.

C. Object-oriented Source Code Model

The Vsound source code meta-model was inspired by the FAMIX [39] information exchange meta-model. The source code meta-model (*cf.* Figure 2) displays the main source code elements and their relations. Mainly, the reader gets enough information if he considers the main type entities that construct an object-oriented system. These are package, class, interface, attribute, method, and the relations between them, such as inheritance, access and invocation. The OO source code model shows structural source code entities such as: packages, classes and methods. In addition, this model represents explicitly

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

information such as: a class inherits from another class (*i.e.*, inheritance), a method accesses attributes (*i.e.*, access) and a method invokes other methods (*i.e.*, invocation). These abstractions are very important and needed for reengineering tasks such as: dependency analysis [40].

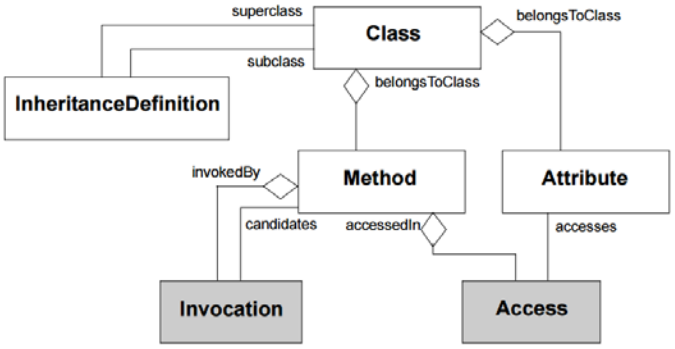


Figure 2. The source code meta-model [39].

D. An Illustrative Example

As an illustrative example, this paper considers the drawing shapes software[3] (*cf.* Figure 3). This software allows a user to draw three different kinds of shapes. The drawing shapes software allows user to draw lines, rectangles and ovals and choose the color of the drawn shape (*cf.* Figure 3). This example used to better explain some parts of this paper. However, Vsound approach only uses the source code of software as input of the documentation process and thus do not know the code dependencies or software metrics in advance.

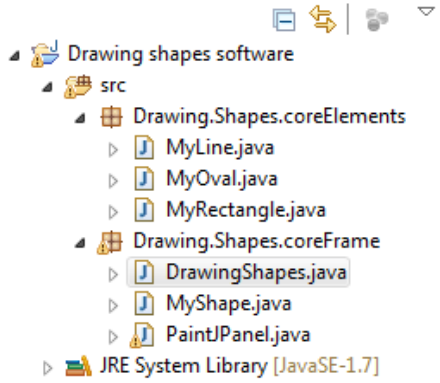


Figure 3. The drawing shapes software.

IV. The Object-Oriented Software Documentation Process Step By Step

This section describes the OO software documentation process step by step. According to Vsound, the approach identifies the software documentation in four steps as detailed in the following.

A. Extracting the Source Code of Software via Static Code Analysis

Static code analysis is the analysis of computer software that is performed without actually executing programs built from that software [4]. While, analyzing the actions performed by a program while it is being executed is called dynamic analysis [5]. Vsound approach accepts as input the source code of software. Then, the proposed approach generates an XML file based on the static code analysis[4]. The mined XML file contains structural information between OO source code elements (*e.g.*, draw method is inside the MyLine class). It also contains structural dependencies between source code elements (*e.g.*, inheritance, attribute access and method invocation). The Eclipse Java Development Tools (JDT) and the Eclipse Abstract Syntax Tree (AST) can be used to access, modify and read the elements of a Java program [37]. ASTs are broadly used in several fields of software engineering. AST is used as a representation of source code [25]. The extracted XML file contains all information needed to document the software by visualizing its code as a set of graphs. Figure 4 shows the mined XML file from the source code of drawing shapes software.

```
▼<Project ProjectName="Drawing shapes software" LinesOfCode="1500">
  ▼<Packages>
    ▶<Package PackageName="Drawing">...</Package>
    ▶<Package PackageName="Drawing.Shapes">...</Package>
    ▼<Package PackageName="Drawing.Shapes.coreElements">
      ▼<Classes>
        ▼<Class ClassName="MyLine" classAccessLevel="public">
          <SuperInterfaces/>
          <Attributes/>
          ▼<Methods>
            ▼<Method MethodName="MyLine" MethodAccessLevel="public">
              ▶<Parameters NumberOfParameters="5">...</Parameters>
              <LocalVariables/>
              ▶<AttributeAccesses>...</AttributeAccesses>
              <MethodInvocations/>
              <MethodExceptions/>
            </Method>
            ▼<Method MethodName="draw" MethodAccessLevel="public">
              ▶<Parameters NumberOfParameters="1">...</Parameters>
              <LocalVariables/>
              ▶<AttributeAccesses>...</AttributeAccesses>
              ▶<MethodInvocations>...</MethodInvocations>
              <MethodExceptions/>
            </Method>
          </Methods>
        </Class>
        ▶<Class ClassName="MyOval">...</Class>
        ▶<Class ClassName="MyRectangle">...</Class>
      </Classes>
    </Package>
    ▶<Package PackageName="Drawing.Shapes.coreFrame">...</Package>
  </Packages>
</Project>
```

Figure 4. The extracted XML file of drawing shapes software.

B. Identifying the Package Document

Vsound approach extracts several documents based on the software source code. These documents cover all granularity levels of the source code. To provide a better understanding of existing software, it is impossible to gather all information in one graph. Vsound approach provides one document (*i.e.*, graph) for every granularity level (*i.e.*, package, class, attribute, method). The package document aims to provide specific information about software. The package document contains

[3] Source code: http://code.google.com/p/drawing-shapes-software/

[4] Source code: https://code.google.com/p/abstract-syntax-tree-vsound/

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

information about software packages and quantitative information about the software (*cf.* Figure 5).

Figure 5 shows the mined package document from drawing shapes software. This document provides information about software packages. As an example, from the graph in Figure 5, drawing shapes software consists of two packages, each one contains three classes. In addition, this document provides information about software metrics such as Lines of Code (LoC), Number of Packages (NoP), Number of Classes (NoC), Number of Attributes (NoA) and Number of Methods (NoM). As an example, from the graph in Figure 5, drawing shapes software consists of 29 methods and 14 attributes.

In order to understand the software source code, the package document is very useful. The goal of this document is to give a general view about software system and present the size of the software (*i.e.*, large, medium or small system). Vsound approach applies to different sizes of software systems. The different complexity levels show the scalability of Vsound approach to dealing with such systems.

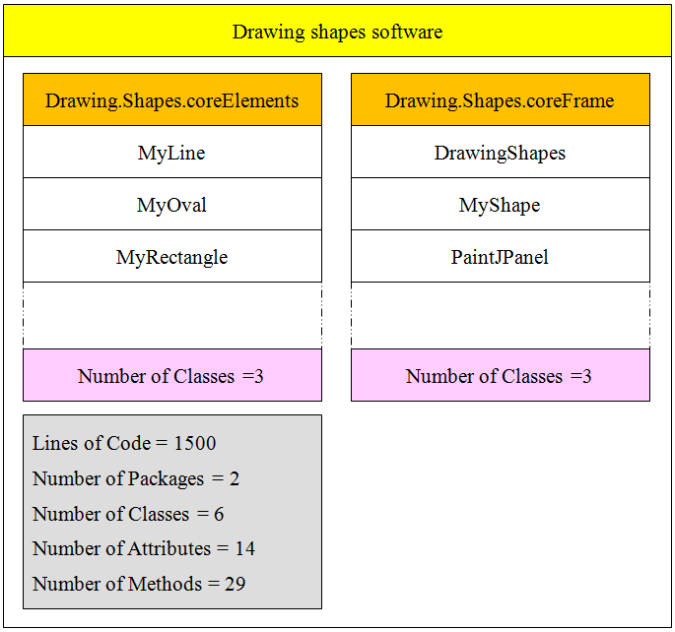


Figure 5. The package document mined from drawing shapes software.

C. *Identifying the Class Document*

The class document displays information about software classes. This information is very helpful toward understanding the software. Vsound approach identifies three documents belong to the class document category. The first document represents the class information document (*cf.* Figure 6). This document shows the number of classes per package. It also presents information about each class in the package. The class information document consists of the class name, super class name, is an interface, super interface name, number of attributes and number of methods. As an example, the class MyShape in Figure 6 consists of five attributes and 12 methods.

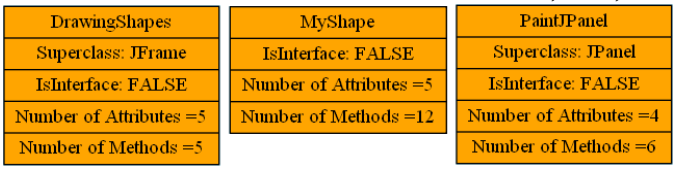


Figure 6. Part of the class information document mined from drawing shapes software.

The second document represents the class dependency document (*cf.* Figure 7). This document shows the main relations between software classes (*i.e.*, Inheritance relation). As an example, the classes MyLine, MyOval and MyRectangle in Figure 7 have a super class called MyShape.

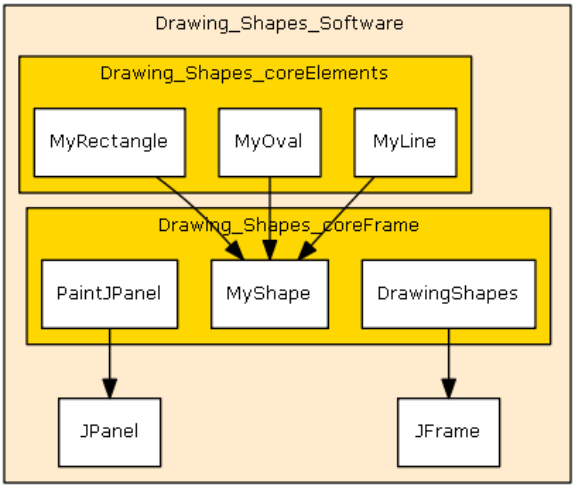


Figure 7. The class dependency document mined from drawing shapes software.

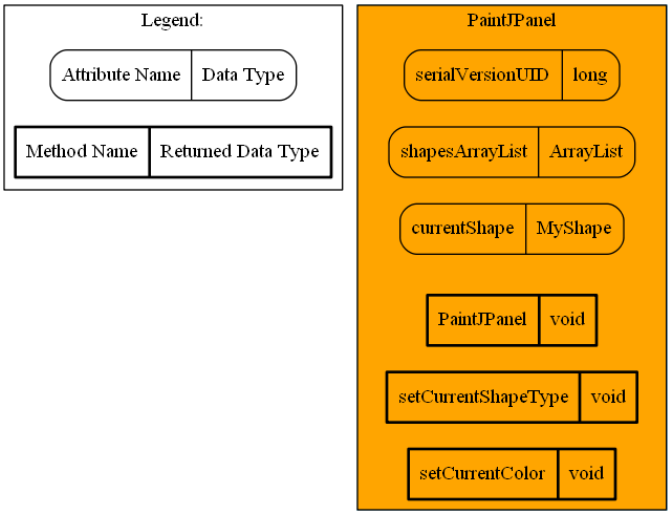


Figure 8. Part of the class content document mined from drawing shapes software.

The third document represents the class content document (cf. Figure 8). This document shows the main content of each class. It also shows the size of class, where the height of class can considered as an indicator of class size. Vsound considers the main elements in the classes. The class content document includes the attribute name and its data type in addition to the

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

method name and its returned data type. As an example, the class PaintJPanel contains currentShape attribute of MyShape type. It also contains setCurrentShapeType method.

D. Identifying the Method Document

The method document shows information about software methods. This information is very useful toward understanding the software system. Vsound approach identifies three documents belong to the method document category. The first document represents the method information document (*cf.* Figure 9). This document provides information about software methods. The method information document contains the following information: the name of a method, the method returned data type, is a static method, number of parameters and the parameter list (*i.e.*, name, data type and order). As an example, the method MyRectangle in Figure 9 consists of 5 parameters.

The second document represents the method content document (*cf.* Figure 10). This document shows the main content of each method. It also shows the size of method, where the height of method can considered as an indicator of method size. Vsound considers the main elements in the method body. The method content document includes the local variable name and its data type, attribute access name and its type and method invocation name and the declared class. As an example, method PaintJPanel in Figure 10 contains addMouseListener which is a method invocation.

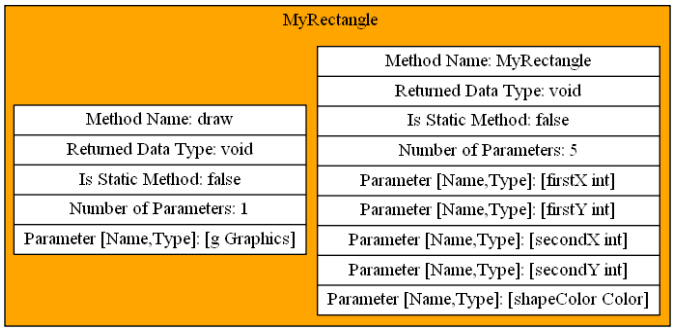


Figure 9. Part of the method information document mined from drawing shapes software.

The third document represents the method dependency document (*cf.* Figure 11). This document shows the main relations between software methods (*i.e.*, method invocation and attribute access). As an example, the method "draw" that is declared in MyShape class is invoked by the paintComponent method in PaintJPanel class. Also, the currentShape attribute that declared at DrawingShapes class is accessed in paintJPanelMouseDragged method that declared at PaintJPanel class.

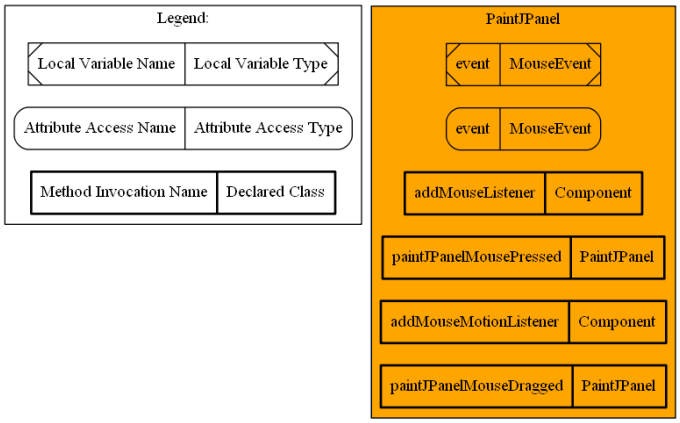


Figure 10. Part of the method content document mined from drawing shapes software.

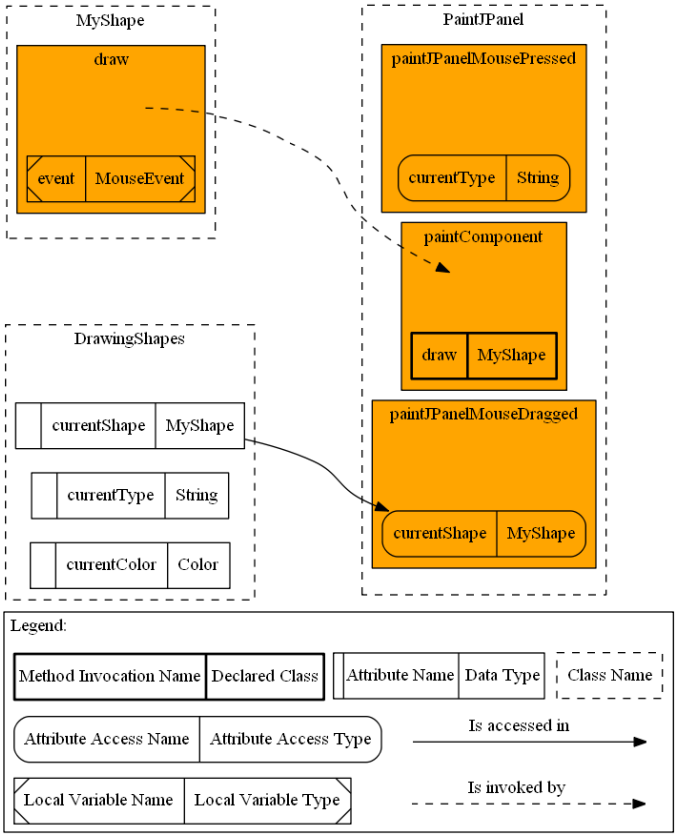


Figure 11. Part of the method dependency document mined from drawing shapes software.

V. EXPERIMENTATION

This section presents the experiment that conducted to validate the Vsound approach. Firstly, this section presents the ArgoUML case study. Next, it presents the evaluation metrics. Then, it also presents detail the architecture and functioning of the Vsound prototype tool and, at last, it presents the software documentation results and threats to validity of Vsound approach.

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

A. Case Study

In addition to the toy drawing shapes example used in this paper, Vsound approach has been tested on other software called ArgoUML. ArgoUML is a widely used open source tool for UML modeling tool. ArgoUML supports the following UML 1.4 diagram types: class diagram, state chart diagram, activity diagram, use case diagram, collaboration diagram, deployment diagram and sequence diagram. The advantage of using the ArgoUML as a case study is that ArgoUML software well documented. In this evaluation, the results are based on the source code of the software that is freely available for downloading in the case study website[5]. ArgoUML runs on any Java platform and is available in ten languages. ArgoUML software is presented in Table 1 characterized by metrics LoC, NoP, NoC, NoA and NoM.

TABLE I. SIZE METRICS FOR ARGOUML SOFTWARE SYSTEM.

Software	LoC	NoP	NoC	NoA	NoM
ArgoUML	120,348	81	1,666	3977	14904

B. Evaluation Measures

In order to evaluate Vsound approach, precision and recall measures are used. In this paper, the precision is the percentage of correctly retrieved links to the total number of retrieved links. The recall is the percentage of correctly retrieved links to the total number of relevant links [36]. In this work, link means: source code element (package, class, attribute, method or local variable) or dependency (inheritance, method invocation or attribute access). All measures (*i.e.*, precision and recall) have values in [0, 1]. If precision is equal to one, every one of retrieved links is relevant. However, relevant links might not be retrieved. If recall is equal to one, all relevant links are retrieved. Nevertheless, some retrieved links might not be relevant. For example, by considering software system contains 95 relations (*i.e.*, inheritance relation). After applying the proposed approach on this software, the result shows that 90 relations are identified correctly (*i.e.*, all retrieved links are relevant). However, five relations are missing. In this case, precision is equal to 1 (*i.e.*, 90/90=1) and recall is equal to 0.94 (*i.e.*, 90/95=0.94).

C. A Simplified Structural View of the Architecture of Vsound

The developed prototype tool [6] of Vsound approach implements the proposed software documentation process. Figure 10 provides an overview of the Vsound tool architecture. It receives as input the software source code and produces as output a set of documents. The XML_Generator module of Vsound produces an XML document which represents the software source code elements and the dependencies between those elements. The tool then starts to generate the software documents by using the DOT_File_Builder component. These documents are serialized as DOT files. DOT is a plain text graph description language. Starting from these DOT files, the tool builds the software document (*i.e.*, graph). For doing so, the Vsound_GUI component of Vsound uses an external library, called Graphviz[7] to produce SVG files containing the documentation of each software artefact (*i.e.*, package, class and method). SVG (Scalable Vector Graphics) is an XML-based file format for describing vector graphics [28].

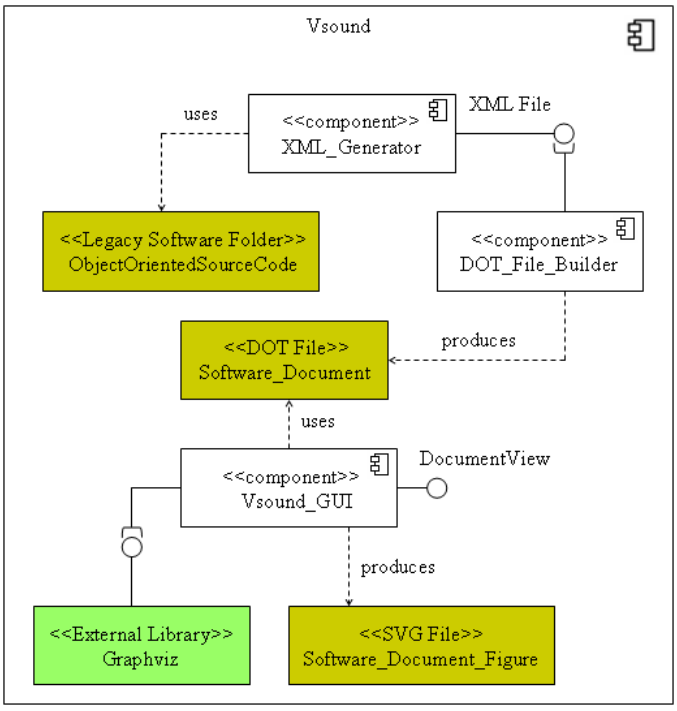


Figure 12. A simplified structural view of the architecture of Vsound.

D. Result

Vsound approach has tested on the ArgoUML case study and obtained promising results. The preliminary evaluation of Vsound shows the significance of this approach. Vsound approach extracted a collection of documents from the source code of ArgoUML software. The first document represents the package document which represents general information about ArgoUML such as: number of packages, number of classes per each package and quantitative information about ArgoUML like LoC and NoM. The second document represents the class document (i.e., the class information document, the class dependency document and the class content document). The identified documents by Vsound provide useful information about software classes. The third document represents the method document (i.e., the method information document, the method content document and the method dependency document). This document shows meaningful information about the software methods. Results show that precision appears to be high for all mined documents from ArgoUML software. This means that all mined documents grouped as software documentation are relevant. Considering the recall metric, recall is also quite high. This means that most source code elements and their dependencies that compose software documentation are mined. Thanks to Vsound approach that identifies software documentation in a novel way.

[5] ArgoUML: http://argouml.tigris.org/
[6] Vsound source code: https://code.google.com/p/fidd/issues/detail?id=2
[7] http://www.graphviz.org/

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

E. Threats to Validity

Vsound approach considers only the Java software systems. This represents a threat to prototype validity that limits Vsound implementation ability to deal only with software systems that are developed based on the Java language. Vsound assumes that source code elements and the dependencies between those elements can be determined statically, such as ArgoUML used in Vsound evaluation. Nevertheless, there exist systems that only behave differently depending on runtime parameters. For such systems, there is a need to extend Vsound to include dynamic analysis techniques. Vsound approach assumes that software documentation can be determined graphically as a set of documents based on the source code. In some cases, there is a need to document software systems by describing their functionalities. This means that Vsound maybe not reliable (or should be improved with other techniques) in all cases to identify software documentation. Documenting software using the names of source code elements and dependencies in its implementation is not always reliable. Thus, it is important to use the comments (*i.e.,* line and block comment) in the source code in order to enhance the documentation process.

VI. Related Work

This section presents the related work relevant to Vsound contribution. It also provides a concise overview of the different approaches and shows the need to propose Vsound approach.

Wettel *et al.,* [11] proposed an approach to visualize object oriented software as a city to solve the navigation problem. CodeCity is a visualization tool that represents the software with a city metaphor. The classes are represented as buildings and the packages as districts. Moreover, some of the visual properties (*i.e.,* width, height, position, and color) of the city artifacts carry information about the software element they represent (*e.g.,* the height of the building represents the number of methods). Their approach does not consider dependencies between source code elements.

Hammad *et al.,* [32] used software visualization techniques to visualize class coupling based on analyzing the source code statically. Other works such as [33] and [34] applied software visualization to model the dynamic behavior of the software by instrumenting the source code in order to monitor the program executions. Moreover, visualization techniques can be applied on software documentation to make it easier and more helpful. For example, work in [35] proposed a visualization framework to visualize bug reports that are saved in software repositories. The proposed framework visualized the status changing for selected bugs, as well as, bug-developer relations by using different shapes and colors.

Al-msie'deen *et al.,* [6] [9] [29] [30] proposed an approach called REVPLINE to identify and document features from the object-oriented source code of a collection of software product variants. The authors presented a new way to document the mined feature implementations in [19]. The proposed approach gives as output for each feature implementation, a name and description based on the use-case name and description [7] [8]. REVPLINE approach aims to document the extracted features from a collection of software variants, while Vsound aims to document OO software as a set of graphs. Hammad *et al.*, [31] presented an approach that focuses on analyzing code changes to automatically detect any OO constraints violation without using graphs as in Vsound approach.

Graham *et al.,* [26] used a solar system metaphor to represent the software source code. The main code entities (*i.e.,* packages and classes) are represented as planets encoding software metrics in the planets' size and color. This visualization represents the software as a virtual galaxy consists of many solar systems. Each solar system represents a package. The central star of each solar system represents the package itself, while the planets, in orbit around it, represent classes within the package. For planets, the blue planets represent classes while the light blue planets represent interfaces. Solar systems are shown as a circular formation. The Solar System metaphor represents packages and classes of software without considering methods and their body. In addition, it focuses only on the inheritance relation between classes without considering the attribute access and method invocation relations.

McBurney and McMillan [12] create summaries of Java methods by using local information (*i.e.,* keywords in the method) and contextual information (*i.e.,* keywords in the most important referenced methods). The method's summaries are generated from elements in the method's signature and body. The approach applied only to the method without considering other artifacts (*e.g.,* package and class). The proposed approach used natural language processing and information retrieval techniques. The type of summary classified under abstract Summary.

Haiduc *et al.,* [27] proposed an approach for summarizing the software source code. The source code summaries consider only the software methods and classes. The proposed approach used information retrieval techniques such as latent semantic indexing and vector space model. The approach extracts the text from the source code of the software and converts it into a corpus (*i.e.,* source code corpus creation). Then, the approach determines the most relevant terms for documents in the corpus and includes them in the summary (*i.e.,* generating source code summaries using text retrieval). Their approach does not exploit structural information from the source code.

VII. Conclusions and Future Work

This paper proposed a new approach for documenting and understanding the software system. The novelty of this approach is that it exploits code dependencies and quantitative information in source code to document OO software in an efficient way by means of a set of documents. The Vsound approach has implemented and evaluated its results on several case studies. The results of this evaluation showed that most of the software systems were documented correctly. Regarding future work, Vsound approach plans to automatically generate descriptive comments that summarize all software packages, classes and methods. It also plans to use the line and block comments in the documentation process.

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

REFERENCES

[1] V. Rajlich and N. Wilde, "The role of concepts in program comprehension," in Proceedings of the 10th International Workshop on Program Comprehension, ser. IWPC '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 271–275.

[2] S. Diehl, Software Visualization - Visualizing the Structure, Behaviour, and Evolution of Software. Springer, 2007.

[3] P. Caserta and O. Zendra, "Visualization of the static aspects of software: A survey," IEEE Trans. Vis. Comput. Graph., vol. 17, no. 7, pp. 913–933, 2011.

[4] Y. Smaragdakis and C. Csallner, "Combining static and dynamic reasoning for bug detection," in Tests and Proofs, First International Conference, TAP 2007, Zurich, Switzerland, February 12-13, 2007. Revised Papers, 2007, pp. 1–16.

[5] M. Egele, T. Scholte, E. Kirda, and C. Kruegel, "A survey on automated dynamic malware-analysis techniques and tools," ACM Comput. Surv., vol. 44, no. 2, p. 6, 2012.

[6] R. Al-Msie'deen, A. Seriai, M. Huchard, C. Urtado, and S. Vauttier, "Mining features from the object-oriented source code of software variants by combining lexical and structural similarity," in IEEE 14th International Conference on Information Reuse & Integration, IRI 2013, San Francisco, CA, USA, August 14-16, 2013, 2013, pp. 586–593.

[7] R. Al-Msie'deen, A. Seriai, M. Huchard, C. Urtado, and S. Vauttier, "Documenting the mined feature implementations from the object-oriented source code of a collection of software product variants," in The 26th International Conference on Software Engineering and Knowledge Engineering, Hyatt Regency, Vancouver, BC, Canada, July 1-3, 2013., 2014, pp. 138–143.

[8] R. Al-Msie'deen, M. Huchard, A. Seriai, C. Urtado, and S. Vauttier, "Automatic documentation of [mined] feature implementations from source code elements and use-case diagrams with the REVPLINE approach," International Journal of Software Engineering and Knowledge Engineering, vol. 24, no. 10, pp. 1413–1438, 2014.

[9] R. Al-Msie'deen, M. Huchard, A. Seriai, C. Urtado, and S. Vauttier, "Reverse engineering feature models from software configurations using formal concept analysis," in Proceedings of the Eleventh International Conference on Concept Lattices and Their Applications, Ko˘sice, Slovakia, October 7-10, 2014., 2014, pp. 95–106.

[10] S. Haiduc, J. Aponte, and A. Marcus, "Supporting program comprehension with source code summarization," in Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering – Volume 2, ser. ICSE '10. New York, NY, USA: ACM, 2010, pp. 223–226.

[11] R. Wettel, M. Lanza, and R. Robbes, "Software systems as cities: A controlled experiment," in Proceedings of the 33rd International Conference on Software Engineering, ser. ICSE '11. New York, NY, USA: ACM, 2011, pp. 551–560.

[12] P. W. McBurney and C. McMillan, "Automatic documentation generation via source code summarization of method context," in Proceedings of the 22Nd International Conference on Program Comprehension, ser. ICPC 2014. New York, NY, USA: ACM, 2014, pp. 279–290.

[13] J. T. Stasko, M. H. Brown, and B. A. Price, Eds., Software Visualization. Cambridge, MA, USA: MIT Press, 1997.

[14] H. A. M¨uller, S. R. Tilley, and K. Wong, "Understanding software systems using reverse engineering technology perspectives from the rigi project," in Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research, October 24-28, 1993, Toronto, Ontario, Canada, 2 Volumes, 1993, pp. 217–226.

[15] K. H. Bennett and V. T. Rajlich, "Software maintenance and evolution: A roadmap," in Proceedings of the Conference on The Future of Software Engineering, ser. ICSE '00. New York, NY, USA: ACM, 2000, pp. 73–87.

[16] E. J. Chikofsky and J. H. Cross II, "Reverse engineering and design recovery: A taxonomy," IEEE Softw., vol. 7, no. 1, pp. 13–17, Jan. 1990.

[17] N. Pereira, M. J. V. Pereira, and P. R. Henriques, "Comment-based concept location over system dependency graphs," in 3rd Symposium on Languages, Applications and Technologies, SLATE 2014, June 19-20, 2014 - Braganc¸a, Portugal, 2014, pp. 51–58.

[18] C. W. Krueger, "Software reuse," ACM Comput. Surv., vol. 24, no. 2, pp. 131–183, Jun. 1992.

[19] R. Al-Msie'Deen, "Reverse engineering feature models from software variants to build software product lines: REVPLINE approach," Theses, Universit´e Montpellier II - Sciences et Techniques du Languedoc, Jun. 2014.

[20] N. H. Madhavji, J. Fernandez-Ramil, and D. Perry, Software Evolution and Feedback: Theory and Practice. John Wiley & Sons, 2006.

[21] S. A. Bohner, "Impact analysis in the software change process: a year 2000 perspective," in 1996 International Conference on Software Maintenance (ICSM '96), 4-8 November 1996, Monterey, CA, USA, Proceedings, 1996, pp. 42–51.

[22] A. Forward and T. C. Lethbridge, "The relevance of software documentation, tools and technologies: A survey," in Proceedings of the 2002 ACM Symposium on Document Engineering, ser. DocEng '02. New York, NY, USA: ACM, 2002, pp. 26–33.

[23] R. Koschke, "Software visualization in software maintenance, reverse engineering, and re-engineering: A research survey," Journal of Software Maintenance, vol. 15, no. 2, pp. 87–109, Mar. 2003.

[24] J. I. Maletic, A. Marcus, G. Dunlap, and J. Leigh, "Visualizing object-oriented software in virtual reality," in 9th International Workshop on Program Comprehension (IWPC 2001), 12-13 May 2001, Toronto, Canada, 2001, pp. 26–35.

[25] G. Rakic and Z. Budimac, "Introducing enriched concrete syntax trees," CoRR, vol. abs/1310.0802, 2013.

[26] H. Graham, H. Y. Yang, and R. Berrigan, "A solar system metaphor for 3d visualisation of object oriented software metrics," in Australasian Symposium on Information Visualisation, InVis.au, Christchurch, New Zealand, 23-24 January 2004, 2004, pp. 53–59.

[27] S. Haiduc, J. Aponte, L. Moreno, and A. Marcus, "On the use of automated text summarization techniques for summarizing source code," in 17th Working Conference on Reverse Engineering, WCRE 2010, 13-16 October 2010, Beverly, MA, USA, 2010, pp. 35–44.

[28] N. A. Aboud, G. Ar´evalo, J. Falleri, M. Huchard, C. Tibermacine, C. Urtado, and S. Vauttier, "Automated architectural component classification using concept lattices," in Joint Working IEEE/IFIP Conference on Software Architecture 2009 and European Conference on Software Architecture 2009, WICSA/ECSA 2009, Cambridge, UK, 14-17 September 2009, 2009, pp. 21–30.

[29] R. Al-Msie'deen, A. Seriai, M. Huchard, C. Urtado, S. Vauttier, and H. E. Salman, "Mining features from the object-oriented source code of a collection of software variants using formal concept analysis and latent semantic indexing," in The 25th International Conference on Software Engineering and Knowledge Engineering, Boston, MA, USA, June 27-29, 2013., 2013, pp. 244–249.

[30] R. Al-Msie'deen, A. Seriai, M. Huchard, C. Urtado, S. Vauttier, and H. E. Salman, "Feature location in a collection of software product variants using formal concept analysis," in Safe and Secure Software Reuse - 13th International Conference on Software Reuse, ICSR 2013, Pisa, Italy, June 18-20. Proceedings, ser. Lecture Notes in Computer Science, J. M. Favaro and M. Morisio, Eds., vol. 7925. Springer, 2013, pp. 302–307.

[31] M. Hammad, M. Hammad, and M. Bsoul, "An approach to automatically enforce object-oriented constraints," International Journal of Computer Applications in Technology, vol. 49, no. 1, pp. 50–59, 2014.

[32] M. Hammad and A. Rawashdeh, "A framework to measure and visualize class coupling," International Journal of Software Engineering and Its Applications, vol. 8, no. 4, pp. 137–146, 2014.

[33] M. Hammad and J. Cook, "Lightweight deployable software monitoring for sensor networks," in The 18th International Conference on Computer Communications and Networks, San Francisco, California, August 3-6, 2009., 2009, pp. 1–6.

[34] M. Hammad and J. Cook, “Lightweight monitoring of sensor software,” in The 2009 ACM Symposium on Applied Computing, Honolulu, Hawaii, USA, March 9-12, 2009., 2009, pp. 2180-2185.

[35] M. Hammad, S. Abufakher, and M. Hammad, “Visualization approach for bug reports in software systems,” International Journal of Software Engineering and Its Applications, vol. 8, no. 10, pp. 37–46, 2014.

[36] T. F. Bissyandé, F. Thung, S. Wang, D. Lo, L. Jiang, and L. Réveillère, “Empirical evaluation of bug linking,” in The 17th European Conference on Software Maintenance and Reengineering, Genova, Italy, March 5-8, 2013., 2013, pp. 89–98.

[37] J. L. Overbey, “Immutable source-mapped abstract syntax tree: a design pattern for refactoring engine APIs,” in The 20th Conference on Pattern Languages of Programs, Monticello, Illinois, 2013, pp. 1–8.

[38] J. S. Poulin, J. M. Caruso, and D. R. Hancock, “The business case for software reuse,” IBM Systems Journal, vol. 32, no. 4, pp. 567–594, 1993.

[39] S. Tichelaar, S. Ducasse, and S. Demeyer, “FAMIX: Exchange experiences with CDIF and XMI,” in The ICSE 2000 Workshop on Standard Exchange Format, WoSEF 2000, 2000, pp. 1–7.

[40] S. Tichelaar, S. Ducasse, S. Demeyer, and O. Niestrasz, “A meta-model for language-independent refactoring,” in The International Symposium on Principles of Software Evolution, 2000, pp. 157–167.

[41] A. Al-Shamailh, R. AL-msie’deen, and A. Alsarhan, “Comparison between the rules of data storage tools," International Journal of Database Theory and Application, vol. 8, no. 1, pp. 129–136, 2015.

AUTHORS PROFILE

Ra’Fat Al-Msie’Deen is a lecturer at the Information Technology department in Mu’tah University, Al Karak - Jordan. He received his PhD. in Computer Science from University of Montpellier 2, France in 2014. He received his Masters degree in Information Technology from University Utara Malaysia, Malaysia in 2009. He got his B.Sc. in Computer Science from Al - Hussein Bin Talal University, Jordan in 2007. His research interest is software engineering with focusing on software documentation, software visualization, software maintenance, software evolution, reverse engineering, reuse, re-engineering, software product line and formal concept analysis.

The knowledge management strategies used as a tool within and through strategic consulting firms to increase the organisational performance

Fatimetou Zahra Mohamed Mahmoud, Ahmed Mohamed Mahmoud
Faculty of Information and Communication Technology
IIUM
Kuala Lumpur, Malaysia

Jamaludin Ibrahim
Faculty of Information and Communication Technology
IIUM
Kuala Lumpur, Malaysia

Abstract: As a matter of fact, the rising awareness in organisations to not focus only in the physical assets in the organisation but also in other paramount assets such as knowledge was the launching point for organisations to go to the implementation of knowledge management and creating knowledge management strategies codified or personalized in order to overcome issues and plan for its future sustainability and development. Actually, the implementation of knowledge management strategy is the strategic consulting company internal asset and external mean to get value. Although, after the implementation of knowledge management strategy in some companies they face the problem of the low revenue and efficiency of the strategy implemented. That's why the combination of the codification and personalization make the organisation able to create, share, save and re-use knowledge and exchange it in the organisation using technology or informal communication networks in order to raise the innovation in the work environment and the creation of new ideas and services. Also, this knowledge stored and codified especially about the organization customers can be considered as the basic knowledge to take the strategic decisions and by KM the organisation will be able to extract, understand and well use this knowledge in order to strengthen the relation with its customers and allow them to become the organisation's partners and advisors. Indeed, that's mean that the organisation strategy must be renewable and in corporation with the overall business model and focus on the core competence in the organisation and the management of the customer knowledge and interactions, this will influence positively and upgrade the organisational performance in any organisation.

Key words: Knowledge management, Strategy, Consulting firms, organisational performance

I-Introduction

Nowadays, there is a significant awareness in all the modern organisations and it start to recognise and know the importance of considering the knowledge as a valuable asset in the organisation that can be managed. [1][2] Actually, knowledge can be divided into two types the explicit knowledge which is the documented and codified knowledge and in this kind of knowledge information technology play an important role to share and save this knowledge by the use of systems and the latest technology in the organisation. Furthermore, the second type is the tacit knowledge which is the knowledge that come by experience in work, intuition and maybe we do not know that we have it or how to explain it, this kind of knowledge can be shared by the interaction and the well communication environment between workers in the organisation. [3][4] From this understanding of knowledge we can say that in a learner and smart organisation, knowledge management is considered as the combination of people, processes, tools and technology to acquire, learn, create, renovate and share knowledge in order to optimise the employment and the well use of this knowledge in the organisation in order to increase the organisational performance and ensuring it's sustainability. [1][5][6][7][8]

In fact, the consulting firms had the precedence to invest in knowledge management. Furthermore,

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

there are different strategies to involve the knowledge management within the consulting companies or any other organisation. The strategies are the codification strategy based on computers and databases to save the knowledge and second strategy is the personalization strategy based on communication among people in work. [9][10]Indeed, any organisation must know and have a clear vision why it need to develop a knowledge management strategy what is the goal of this implementation, in what it can help the organisation to be developed and it should be aligned with the organisation overall strategy. In this situation the consulting firms play an important role to help the organisation to identify clearly what are the weaknesses, the issues in the organisation that need to be resolved and to focus on helping the organisation to keep and attract customers by implementing a knowledge management strategy which will be related to the global long-term vision of the organisation and supply a framework that will help the organisation to resolve and overcome those issues and increasing the business and organisational performance. [2][9][11]

II-Literature review

In fact, consulting firms are the more concerned by knowledge because it represent an asset, resource, product or even a service which help the consulting firm to create value and benefit, that's why they need a strategy to manage this knowledge and be able to sale it as their business. The authors Sven, Dirk, Dietrich and Chiara did an analysis on the correlation between the knowledge management strategy and the business model of the organisation. Actually, the management consulting firms provide advice to the organisations relating to the strategy, operations or information technology based on knowledge to add value to the organisation and to improve their performance. Commonly the consulting firms forward one of two business models, the first one is the creation of high customised solutions by the firm for a significant and unique issues by providing original and analytical advices based on experience and tacit knowledge ,this kind of strategic consulting firms business model focus will be to preserve a high profit for the organisation, whilst the second business model followed by the consulting firms is the providing of a high standardised products and services due to their dealing with the same issues thus they reuse their existing modules and they continue to create new modules and pieces in order to generate great revenues.[12]

In the other hand in consulting firms the knowledge management strategy follow specific goals, techniques and technology and the generation, distribution, maintenance of knowledge in each process is controlled centrally which mean the knowledge will be codified, stored in data bases and documented or de- centrally which mean a personalized knowledge attached to the person who acquire it. Then the authors consider that the firms which has a standardized business model should focus on the central knowledge management strategy, whereas the consulting firms that provide the customised solutions should focus on the de-central knowledge management strategy because in this case the codified knowledge will have a limited value for the organisation and they need to utilize the tacit knowledge and experience to innovate solutions. Furthermore, this analysis was proved after studying the business model and knowledge management strategy in the case of leading companies each one in a specific sector such as Mckinsey, Accenture, Price water house Coopers, and prognos AG. [12]

Another study analyse the two types of strategies the codification and personalization from a marketing perspective to optimize the efficiency of knowledge reuse. Actually, one of the issues is that many organisation suffer due to the depressed returns from the investment in knowledge management. Thus, to significantly increase the efficiency of transferring knowledge among consumers and producers, we should first know that the inefficiency of knowledge transfer is due to the diverse priorities and agendas of the producers and consumers pending the knowledge exchange and share. Indeed, the authors after developing a model to help in maximizing the efficiency of knowledge reuse and transfer they found that the two strategies codification and personalization should be combined together to enhance the efficiency of knowledge reuse.[13]

From another point of view, a study and research was done by Zhu Yu, Wang Yan-fei and Lan Hai-lin with the participation of 223 enterprises in China to recognize the relation between the knowledge management strategy, the core competences and the organizational performance. [14]

Actually the study found that, firstly for an organisational future development there should be a coordination between the knowledge management strategy and the core competence in the organisation. It's because the core competence represent the intermediate to create an impact and effect of the knowledge management strategy on the organizational performance and the organisation should pay more attention and focus in the developing of their core competences as well as their strategy to ensure a long- term development and sustainability. Secondly, the researches demonstrate that the greatest impact on human resource competence and efficiency is from the knowledge management strategy and the core competence in the organisation. In addition, the

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

knowledge sharing culture has to be developed in any organisation. Actually, by the implementation of a codified knowledge management strategy the organisation will build a sharing mechanism by the capture, save and reuse of knowledge and it start by the spread, educate and aware employees to better understand what is the sharing culture and how it represent a benefit for the employees and for the organisation. [14]

To resume the study found that the two knowledge management strategies have different impact and influence in the organisation, actually the personalization strategy has a positive impact on the core competence while the codification knowledge management strategy has a positive impact in both the core competence and the organizational performance. [14]

Actually we cannot forget the significance of technology in the improvement and development of any organisation, from this standpoint Barney, Shan and Ray had did an analysis study on a successful case in Singapore to see how in specification the web technologies can improve the organisational performance in dependence with the global organisational environment. In fact, the authors found that the web technologies can play an important role to assist the organisation business strategy and optimizing its performance. Furthermore, to have an effective influence of web technologies in the organisation the organisation should not depend on the technical concession but in the complicated fit among the strategy, the technology and the external environment. Indeed, the equilibrium in the organisation environment help the web technology to improve the organizational performance by simplifying the realization of competitive features using three different mechanisms *"the logics of positioning, leverage, and opportunity"*, reciprocally in the revolution state of the organisation environment *"Web technologies can give rise to performance gains by supporting the attainment of legitimacy through two distinct mechanisms: the logics of optimality and social congruence"*.[15]
In the same context, another research has shown how the combination between the technology and the knowledge management can be a tool to increase the performance and the profit of the organisation which is the use of customer relationship management (CRM) in integration with knowledge management. Actually CRM as an approach based on strategies and technology help the organisation to ameliorate its business relationship with customers by collecting information about their customers through various points of contact among the organisation and its customers such as the social media, organisation's website, email, call centre and different marketing tools. Thus to achieve this the organisations use different software's to store all the customer information into a single database to record the customer interactions and the automation of workflow processes. But, this information stored in big database cannot be significant and play a role to maintain permanent customers and increase the production and long-term profit in the organisation if this information is not well managed, organized, connected and distributed and here the knowledge management can play a role to extract the meaningful knowledge from those information to transform it into valuable information that can be analysed to attract customer and improving the business performance in the organisation. Indeed, the combination of KM and CRM can help to use the knowledge for, from and about customers by the experts in the organisation in order to attain the organization goals and optimizing its business and organizational performance.[16][17]

One of the successful organisations IBM traditionally was known as a company that have a profound experience in information technology and in their old business model the company was depending on the sale of hardware to get revenues. However, IBM has recognised that the organisations start to give more attention to the strategic value in information technology tools and re-engineering projects and those projects should be linked with the overall business strategy of the organisation. For this, in order to keep up and react with the business environment IBM has created the IBM global services business unit, then the largest acquisition on the IBM history was done by acquiring the Price Waterhouse Coopers (PwC) which is a consulting firm that provide information technology services and IT management consulting. Furthermore, IBM had create a new unit which is IBM business consulting services to combine the global services business unit and the PWC consulting. [18]

Actually, with a very large number of consultants in different countries IBM has become one of the largest consulting services organisations. In the other hand, IBM consulting has involve the intellectual capital management (ICM) in order to formalize the knowledge management over the IBM services and industries and give more attention to the acquire, creation, sharing, using and transferring knowledge in order to ensure the development of the organisation. [18]

In fact, the knowledge management strategy of IBM included the linking between the strategy and the intellectual capital of the company, the creation of a culture based on knowledge, creating processes and infrastructure that help to crate and share knowledge, using technology for sharing knowledge, and the measurement of the intellectual asset sharing effectiveness. Also, one of the IBM priority was to raise the capability of the

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

consultants group to share their knowledge and be able to merge it rapidly in response to the client need by the creation of informal networks as a way to maximize the internal team work liberty to act and share knowledge. Thus, the knowledge management in IBM was based on the society network and what differentiates the process of knowledge management in IBM is the use of the SNA (social network analysis) as a tool to analyse and check the knowledge network characteristics and as a change management initiative. [18]

Furthermore, IBM had realize that the real value come from crating and sharing knowledge, and actually the valuable knowledge of the company come from the heads of the talented employees. Indeed, IBM in order to ensure its sustainability an maintain a reputable brand in consulting need to merge the two knowledge management strategies the top down centralized and the bottom-up decentralized approaches in order to build trust in the work environment, exchanging talents internally as the case of Mckinsey consulting firm where the consultants nominate themselves for a specific project and then they make the manager bid for them, actually this exchange mechanism help the best employees to distinguish their experience and knowledge. [18]

As a matter of fact, IBM is known as an ongoing strategic renewal company as mentioned by the authors Edna and Tuvya. The success of IBM all over the years was based on a strategic conjunction of innovation in business management and technology. Actually, in the level of technological innovation IBM is known as one of the company that have the largest research and development set and the success of this set is due to the management of employees and their knowledge. Furthermore, due to the research and development set, IBM invented as a business model the employment of personal computers, then the mobile computing and the think Pad, after this IBM has made a great transformation by focusing on providing consulting to clients. Thus, IBM was a company that make many transformations according to the market changing and needs. [19]

In the level of business management innovation, IBM had a long term vision that was the cause of its success today as a company that supply services and solutions, hardware, software with a very profound knowledge in the different fields to which the company supply its services. Also, IBM initializes the internal innovation in the company by focusing and giving importance to processes such as the global brainstorming. Actually, innovation is one of the important sources which is extracted from the internal communication network in the organisation that regroup a very large number of customers , employees and business partners in order to share the new services, ides, conducting debates and much more.[19]

III- Literature findings

The general awareness in organisations from different fields and sectors on the importance to consider knowledge as a valuable asset, give raise to the important role of strategic consulting firms which are themselves need the knowledge internally as an asset and resource, to create product or a service in order to get value and benefit, or externally to sale it as their business by implementing strategies to manage knowledge in other organisation from different fields.

Actually, there is two types of knowledge management strategies used within and through the consulting companies which are the codification and personalization strategies. Indeed the implementation and the success of the knowledge management is by the combination of this two strategies which mean that the knowledge should be saved and codified to simplify its reuse and in the same time it should be exchanged through the personalized strategy. In addition, this combination of two strategies should be related to other components of the organisation such as the technology role, the business model role, and the importance of core competences and the creation of a sharing culture, the importance of the intellectual capital and its assessment and the rapid adaption to the market change.

As a result, the combination and the alignment of the components mentioned before with the knowledge management strategies can play an important role to increase the organizational and business performance significantly and ensure the development and the sustainability and help it to keep a high level of their customer satisfaction of the organisation such as the successful case of IBM Company.

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

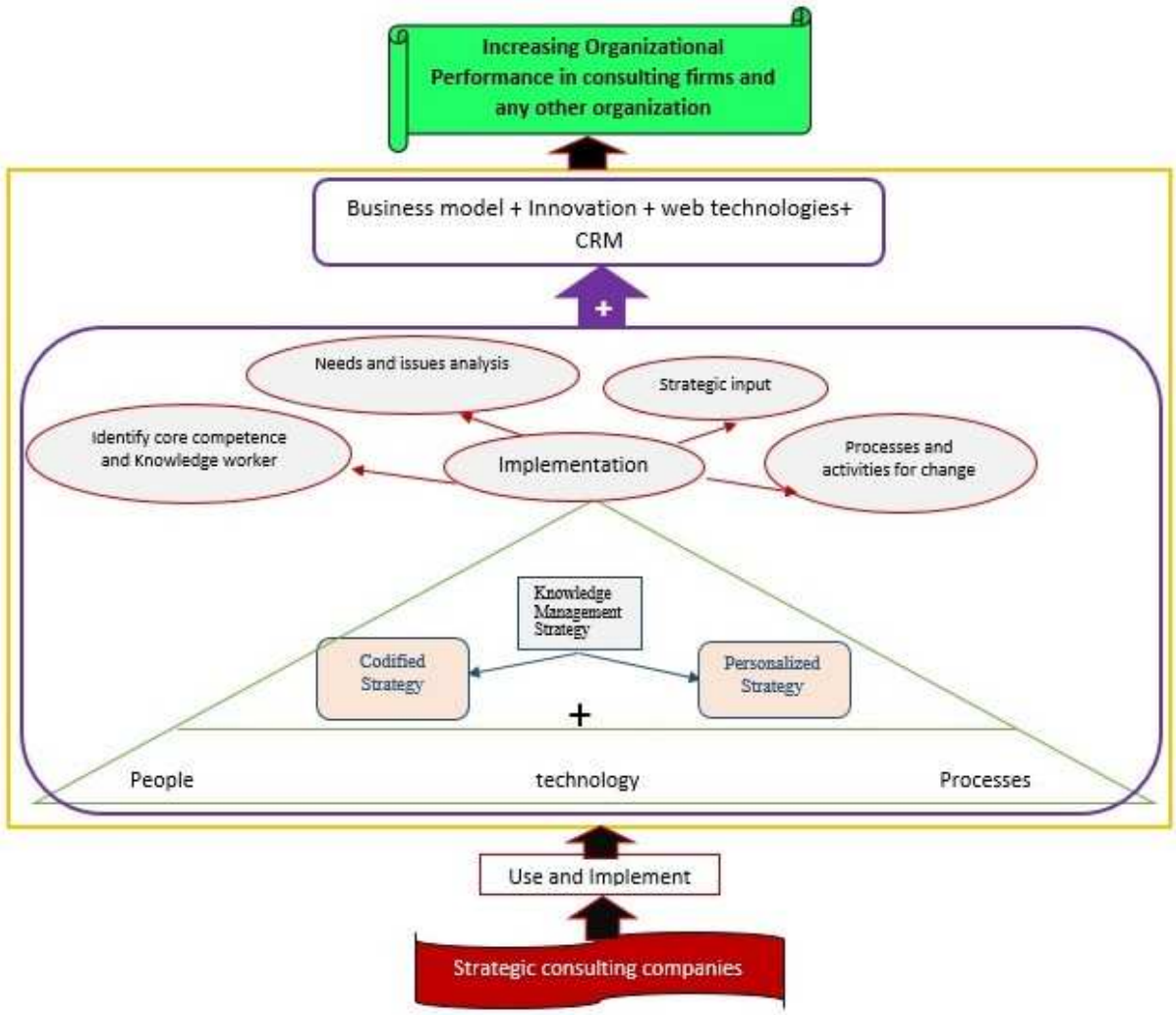


Figure 1

In the figure above we see how the strategic consulting firms use and implement the knowledge management strategies in addition and alignment with the business model, by focusing in the innovation in the organisation and the use of technologies such as the web technology and the customer relationship management systems. All of these components together lead to the increase in organisational performance in the consulting firm itself and in the other organisations that the consulting firms help to improve their situation and resolve their issues.

IV-Conclusion

In short,the knowledge management strategy help the organisation to analyse its need, issues, and innovate solutions and plans for the development of the company and its sustainability by reachinga high level of satisfaction from its customers and increasing the interactions with them by using technology and systems or by the direct communications and interviews which make the customers partners in defining the organisation future, and this is the goal of consulting firms in the implementation of a knowledge management strategy in an organisation. Actually, the successful implementation of knowledge management strategy in consulting firms or by consulting firm to another company play a major role in contribution with other components such as the business model to have a significant potential increase of the organisational performance in the organisation.

V-References

[1]. Edna Pasher, TuvyaRonen.*"The complete guide to knowledge management"*.Wiley. February 8, 2011

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

[2] Paul S. Chinowsky, School of Civil and Environmental Engineering Paul S Chinowsky, Anthony D. Songer.*"Organization Management in Construction"*. First edition. Jan 12, 2011. 216 pages.

[3]. Alaaeddin M. Abed, Md. Mahfuzul Islam, Khalid Helal, Mohamed Abdirahman, Abdirizak Mohamed, Dr.Jamaludin Ibrahim*." The Manager and User's Role in IT Use as a KM Tool"*. International Journal of Scientific and Research Publications, Volume 4, Issue 12, December 2014

[4]. KimizDalkir, McGill University. *"Knowledge management in theory and practice"*. Elsevier Butterworth–Heinemann.

[5]. Gul Muhammad, Jamaludin Ibrahim, Zeeshan Bhatti*, Ahmad Waqas*." Business Intelligence as a Knowledge Management Tool in Providing Financial Consultancy Services"*. American Journal of Information Systems, 2014, Vol. 2, No. 2, 26-32

[6]. DanijelaJelenic. Faculty of Economics, university of Nis, republic of Serbia*." The importance of knowledge management in organizations – with emphasis on the balanced Scorecard learning and growth perspective"*. Management knowledge and learning. International conference 2011.

[7]. Kamran Nazari, Department of Business Management, Payam Noor University, Kermanshah, Iran. MostafaEmami, Young Researchers Club, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran*." Knowledge Management: FROM THEORY TO PRACTICE"*. Australian Journal of Business and Management Research. Vol.1 No.11 [22-30] | February-2012

[8]. Gillingham, Helen, and Bob Roberts. *"Implementing knowledge management: a practical approach."* Journal of Knowledge Management Practice 7.1 (2006).

[9]. Morten T. Hansen, Nitin Nohria, Thomas J. Tierney. *"What's your strategy for managing knowledge?*". Harvard Business review. March-April 1999

[10]. Karsten Jahn. *"Knowledge management strategies*". January 12, 2013. Retrieved from: http://studies.k-jahn.de/2013/01/12/knowledge-management-strategies/

[11]. Manjeet Sharma. *"Knowledge management: principle and strategy*". Published by: Sai Om Journal of Commerce & Management. Volume 1, Issue 10 (October, 2014)

[12]. Sven Grolik, Dirk Kalmring, Dietrich Lehner, Chiara Frigerio." *Analysis of Interrelations between Business Models and Knowledge Management Strategies in Consulting Firms"*.KM Strategies in Consulting Firms

[13]. Kah-Hin Chai, James Nebus*." Personalization or Codification? A Marketing Perspective to Optimize Knowledge Reuse Efficiency"*. IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, VOL. 59, NO. 1, FEBRUARY 2012

[14]. ZHU Yu, WANG Yan-fei, LAN Hai-lin*." Empirical Study on Interaction among Knowledge Management Strategy, Core Competence and Organizational Performance"*. Collage of Business Administration, South China University of Technology, P.R.China, 150001

[15]. Barney C. C. Tan, Shan L. Pan, Ray Hackney*." The Strategic Implications of Web Technologies: A Process Model of How Web Technologies Enhance Organizational Performance"*. IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, VOL. 57, NO. 2, MAY 2010

[16]. Ali Attafar, Majid Sadidi, Hamideh, Attafar, Arash Shahin. *"The Role of Customer Knowledge Management (CKM) in Improving Organization-Customer Relationship*". Middle-East Journal of Scientific Research 13 (6): 829-835, 2013

[17]. Nour Mohamed Yaghoubi, Seyed Mohamed sadeq, Seyyed Ali, Asghar Afshar Jahanshahi, Khaled Nawaser. *"The role of knowledge management on customer relationship management"*. European Journal of Economics, Finance and administrative sciences, Issue 34(2011)

[18]. Kylie Summerhayes, Steven Luo*." Managing Knowledge in Professional Consultancy"*. Monash Business Review Volume 2 Issue 3-November, 2006

[19]. Edna Pasher, Tuvya Ronen. *"The complete guide to knowledge management"*.Wiley. February 8, 2011. (Pages: 42, 43, 44)

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

Discrete Fourier Transform Attack on WG-7 Cipher

Rashid Zulfiqar [#1], Mehreen Afzal [#2]

[#] *College of Telecommunication(MCS)*
National University of Sciences and Technology
Islamabad, Pakistan
[1] rashid.is-10@mcs.edu.pk

[2] mehrenafzal@mcs.edu.pk

***Abstract*—This paper presents application of Discrete Fourier Transform(DFT) attack on Stream Cipher Welch Gong(WG)-7. WG-7 is a lightweight, hardware oriented stream cipher that uses a word oriented Linear feed back shift register(LFSR) and a nonlinear WG transformation that acts on the LFSR output word. The cipher has been designed to work in a resource constrained environment by Yiyuan Luo, Qi Chai, Guang Gong and Xuejia Lai as variant of original WG Cipher. This paper aims at faster recovery of keystreams than predicted complexity of the DFT attack by the designers. The proposed DFT attack recovers the keystream with a time complexity of at the most 2^{22} and keybits complexity of at the most 2^{19} by employing the annihilator in the structure of the cipher. The proposed attack is even efficient than Algebraic attack and Fast Algebraic attack on the cipher.**

***Index Terms*—WG-7, Discrete Fourier Transform Attack, Key Recovery Attack, Discrete Fourier Spectra Attack.**

I. Introduction

Ronjom and Helleseth [1] introduced the New Attack on the Filter Generator, to recover the initial state of a filtering sequence generator. The attack is efficient than the Fast Algebraic Attack(FAA), but needs more keybits as compared to FAA. Ronjom and Helleseth [2] extended the same attack [1] to the case of filter generators over finite fields F_{2^n}, consisting of a primitive feedback polynomial over F_{2^n} generating a sequence that is filtered through a nonlinear Boolean function. Also results were presented on attacking WG Cipher proposed by Yasir et al. [3]. Ronjom et al. [4] made the linear subspace attack application generic by forming a system of linear equations over F_{2^n} instead of F_2 utilizing filtered sequences with their trace representation. Gong [5] showed a fast computation of DFT using the selective DFT algorithm resulting in simplifying selective DFT attack. Gong et al. [6] introduced the new variation known as the fast selective DFT attack, which is strongly related to the FAA and algebraic attack, however, fast selective DFT attack is more efficient than FAA when the captured number of consecutive bits of the cipher is less than the linear complexity of the sequence. DFT attack has been further improved by Wang et. al. [7] as it relaxes the successive keystream sequence requirement and carries out DFT operations in subfield rather than extension field. The attack is applied to a version of Bluetooth encryption algorithm E_0 by Wang et. al. [8]. Wang et. al. [9] also established the fact that DFT attack can recover any keystream by using at most half of the cipher period.

WG-7 [10] is a lightweight stream cipher whose design is mainly inspired by the WG stream cipher [3]. WG-7 is hardware oriented stream cipher that uses a word oriented Linear feed back shift register(LFSR) and a WG transformation. It consists of a 23 stage LFSR over F_{2^7} and a WG linear transformation. The cipher has been designed to work in a resource constrained environment. The authors of the cipher have analyzed the cipher design and concluded that this cipher is secure against DFT attack. However, this evaluation are not in-depth especially after the discovery of Annihilator by Muhammad Ali et. al. [11]. The annihilator was discovered to reduce the complexity of the Algebraic attack on WG-7. The same annihilator is used in our proposed attack to depict the possibility of DFT attack against WG-7 cipher. The Fast Selective DFT attack [6] version has been employed which is also referred in literature as Fast Discrete Spectra Attack. The proposed attack reduces the complexity of the attack by placing heavy computation during offline phase. Later on it is shown that proposed DFT attack is more efficient in terms of keybits requirement and online computation when compared with Algebraic and Fast Algebraic attack on the WG-7 Cipher.

The paper is organized as follows. Section II gives a brief description of the WG-7 Cipher, Section III gives a brief description of the Fast Selective DFT attack. Then section IV describes the proposed DFT attack.

II. Description of WG-7

WG-7 [10] is designed to generate up to 2^{24} bits of key stream from an 80-bit key length and an 81-bit initialization vector(IV). The internal state of 161 bits $[s_1, \cdots, s_{161}]$ is divided into one LFSR of length 23, containing 7-bit words each. WG-7 is hardware oriented stream cipher that uses a word oriented Linear feed back shift register(LFSR) and a WG transformation. It consists of a 23-stage LFSR over F_{2^7} and a WG linear transformation. It has the ideal two-level autocorrelation property. The non-linear transformation is defined by the equation 1.The cipher works by loading the secret key and IV to initialize the internal state of the LFSR. Then clocking of the LFSR is carried out 46 times with its non-linear update. After this the cipher generates the required keystreams. The internal state of the cipher is 161 bits which are used to

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

generate 2^{24} bits of keystreams.

$$WG7(x) = Tr(x^3 + x^9 + x^{21} + x^{57} + x^{87}), x \in F_{2^7} \ . \quad (1)$$

Yiyuan et al. [10] have concluded that DFT attack is impossible to be launched against stream cipher WG-7. The attack calculations are based on complexity computations given by Ronjom et al. [2]. The complexity of the attack has been found out to be $2^{25.5}$ key stream bits after a precomputation with a complexity of $O(2^{39.5})$. If the adversary obtains the keybits less than 2^{25}, the adversary has to guess $2^{23.5}$ unknown bits to launch the DFT attack. The cipher designers have concluded that best against the WG-7 is the exhaustive search as adversary cannot obtain 2^{24} consecutive keystream bits.

III. Discrete Fourier Transform Attack

Two DFT attack versions have been defined by Gong et. al. [6]. Selection between the two versions requires the exact linear span of the cipher. If the linear span is high, the selective DFT method cannot be used. However, fast selective DFT Attack can be used in this situation which reduces the linear span by multiplying essentially another sequence having lower linear span than the original sequence. The adversary can launch DFT attack to recover the internal state and then the cipher is clocked backwards to recover the secret key. The fast selective DFT attack is described here in the subsequent section.

A. Fast Selective DFT Attack

The algorithm is description of Algorithm 2 given by Gong et al.in [6]. It recovers the factor β where the number of observed consecutive bits of a filter generator is less than the linear complexity of the sequence. The fast selective method is used to recover the initial state of the cipher. The bits where the output sequence is $\mathbf{1}$ are filtered and their relative index are then used to develop the system of equations. The trace form of the polynomial is used to develop the systems of equations. The trace computation is also carried out to find the elements of subfield F_2 from F_{2^n}. The system of equation is solved using any known methods to recover β.

1) Offline Computation: The off-line computation includes the selection of annihilator function and subsequent conversion to its DFT form(Trace Form) for recovery of keybits in online phase. The major computation is carried out in this phase as a one time measure. First selection of a new sequence $g = g_t$ is carried out with the condition that the original sequence u_t and annihilator sequence g_t product is $\mathbf{0}$. The new sequence also satisfy the following condition that: $H_g < l(u)$, where $H_g = \{k | G_k \neq 0$, k is a coset leader mod H$\}$, H is the complete period of the cipher .The computation of the G_k will be carried out using the method described by Gong et al. [4].

2) Online Computation: In online computation, all k will be replaced with t where keystream are $u_t = 1$. Function v_t will be used to perform the procedure for recovery of β from equation $v_t = \sum_{k \in H_g} Tr(\beta G_k \alpha^k) = 0$. Thereafter the coefficient matrix is computed by replacing $G_k \alpha^k$ with $G_k \alpha^k = x_0 + x_1 \alpha + x_2 \alpha^2 + \cdots + x_{n-1} \alpha^{n-1}$.

$$\begin{pmatrix} v_0 & \cdots & \cdots & v_{n-1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{l(g)-1} & \cdots & \cdots & v_{(n-1)(l(g)-1)} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad (2)$$

The independent system of equations is generated which produces a unique solution. Gaussian Elimination method can be used to solve this system of equations. To recover bits after initialization stage, above system of $l(g) - 1$ linear equations over F_2 with $l(g) - 1$ variables are required to be solved. If there is a $k \in H_g$ with gcd (k,N) = 1, then β will be extracted by $\beta = \beta^{k^{k'}}$, where $k' = k^{-1} \pmod N$.otherwise, $\{\beta^k | k \in H_g\}$ values will be returned. In this Algorithm, the number of required consecutive bits from u_t is l(g).

3) Complexity Calculations: The complexity calculations are same as mentioned in [6]. The preprocessing phase is a one-time effort like Fast Algebraic Attack. The complexity for the online phase of the attack is calculated for case where product of original function sequence and Annihilator sequence is equal to zero.

IV. Proposed Fast Selective DFT Attack on WG-7

This section describes our proposed attack on WG-7 to reduce its complexity on WG-7. The algebraic normal form (ANF) of non-linear transformation acting on last 7 bits of the cipher is given by Mohammad Ali et al. in [11]. The best annihilator found out by Mohammad Ali et al. [11] of non-linear transformation function g is given as equation 3.

$$\begin{aligned} h(y_1, ..., y_7) = {} & 1 + y_1 + y_3 + y_1y_2y_3 + y_4 + y_1y_4 + \\ & y_2y_4 + y_1y_2y_4 + y_3y_4 + y_4y_5 + y_1y_4y_5 + y_3y_4y_5 + \\ & y_6 + y_1y_6 + y_2y_6 + y_1y_2y_6 + y_3y_6 + y_2y_3y_6 + y_2y_3y_7 + \\ & y_4y_7 + y_2y_4y_7 + y_3y_4y_7 + y_3y_5y_7 + y_4y_5y_7 + y_6y_7 + \\ & y_1y_6y_7 + y_2y_6y_7 + y_3y_6y_7 + y_1y_3y_4 + y_2y_3y_4 + y_7 + \\ & + y_3y_7 + y_1y_3y_7 + y_1y_3y_5 \ . \end{aligned} \quad (3)$$

The application of Fast Selective DFT Attack proposed by Gong et. al. [6] on WG-7 cipher cannot achieve considerable reduction in complexity of the attack due to structure of the cipher (WG-7). Another limitation it poses is due to requirement of keystream in succession regardless of $\mathbf{1}$ and $\mathbf{0's}$ for success of the attack. Also the structure of the word oriented cipher hinders in computation of the DFT for the complete period $(2^{161}-1)$ and the primitive polynomial which completely defines this period. This method transforms the structure of LFSR of WG-7 into an equivalent polynomial form of degree 161 before application of the filter function. A primitive polynomial will be derived to define the complete period of the WG-7 cipher before application of filter function. The polynomial derived as a result will be used to calculate

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

DFT form of the annihilator and original function.This devised method achieves significant decrease in complexity than DFT attack showed by Yiyuan et.al. [10] and employs Fast Selective DFT Attack [6]. To further illustrate this procedure, a simulated cipher similar in structure to WG-7 has been used. Two phases of the attack will be utilized to perform this method, offline and online.

A. Offline computation

In offline phase, polynomial form of the cipher LFSR structure will be developed which generates all the elements of the field $GF(2^{161})$. For this purpose, two outputs of the cipher will be taken simultaneously. First output will be generated without the original filter function given as 1 and second with annihilator given as 3. The first output of the LFSR stage of the cipher will generate 2^{161} bits. The bits will be used to calculate the primitive polynomial of the cipher using the well known Berkeley and Mason (BM) algorithm. The polynomial will be primitive in $GF(2^n)$ where n= 161. The second output of the cipher using annihilator from equation 3 will be generated and stored with initial state set as $(0, 0, 0, 0, 0, 0, \alpha^{128})$. The polynomial of $GF(2^n)$ calculated above will be used to calculate DFT coefficients (Trace form) using the annihilator's output. The selective DFT computation method described by Gong et al. [4] will be used to compute DFT. Cyclotomic cosets leaders will be utilized to represent the equivalent Trace form of the annihilator for launching fast selective DFT attack as described by Gong et al. [6]. The major computation of the attack will be carried out in precomputation phase and it is a one time effort.

B. Online Computation

Significant decrease in linear span of the cipher will occur due to use of Annihilator. This also reduces the number of required linear independent equations. The system of equation will be generated where all captured keystream are **1** to recover the initial state of the filter Generator (WG-7) as per method already described in section III. The system of equation will be solved to generate the keybits for recovering the initial state of the cipher. Complexity of the attack is mentioned in the Table I.

C. Application of DFT Attack on Simulated Filter Generator

A filter generator has been introduced to imitate the proposed method. The relevant details are given in succeeding subsections:-

1) Specification of Filter Generator: Filter generator is designed to generate up to $2^{10} - 1$ bits of keystream from an 10 bit key length. The internal state of 10 bits $[s_0, \cdots s_9]$ is divided into one LFSRs of length 2, containing 5 bit word each. The cipher consists of a 2 stage LFSR over $F_(2^5)$ and a linear transformation. The finite field is defined by primitive polynomial $d(x) = x^5 + x^3 + 1$. The characteristic polynomial is primitive over $F_(2^5)$ and is given by $f(x) = x^2 + x + \alpha^7$, where α^7 is root of d(x). The non-linear transformation is defined by equation 4:-

$$u(x_0, \cdots x_4) = x_0 + x_0x_1 + x_0x_2 + x_0x_1x_2 + x_3 + x_1x_3 + x_2x_3 + x_1x_2x_3 + x_4 + x_0x_4 + x_1x_4 + x_0x_1x_4 + x_2x_4 + x_0x_2x_4 + x_1x_2x_4 + x_0x_3x_4 \ . \quad (4)$$

2) Offline Mode Computation: In offline mode, two outputs of the word oriented LFSR are taken. The period of the cipher is $2^{10} - 1 = 1023$, as n=5x2=10. One output is without its output filter function u(x) 4 for the complete period. The other is with the use of annihilator mentioned as equation 5. Both the outputs are stored with a storage requirement of 2^{11}.

$$g(x_0, ..., x_4) = x_2 + x_0x_2 + x_1x_2 + x_0x_3 + x_1x_3 + x_1x_4 + x_3x_4. \quad (5)$$

The output of the LFSR word is XOR in order to convert it from F_{2^5} to F_2. If all the bits can be represented as $[x_0, \cdots x_9]$, then XOR function is performed on the bits as $x_0 + x_1 + x_2 + x_3 + x_4$. Then it is used as input to BM algorithm to generate the primitive polynomial of the cipher mentioned as equation 6:

$$l(x) = x^{10} + x^9 + x^8 + x^4 + x^3 + x^2 + 1; \quad (6)$$

After this step, the Trace form for the annihilator is generated as per algorithm specified in [5]. The same can be reverified by using the output of the cipher with annihilator and calculating DFT for their cyclotomic cosets leaders: [1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 69, 71, 73, 75, 77, 79, 83, 85, 87, 89, 91, 93, 95, 99, 101, 103, 105, 107, 109, 111, 115, 117, 119, 121, 123, 125, 127, 147, 149, 151, 155, 157, 159, 165, 167, 171, 173, 175, 179, 181, 183, 187, 189, 191, 205, 207, 213, 215, 219, 221, 223, 231, 235, 237, 239, 245, 247, 251, 253, 255, 341, 343, 347, 351, 363, 367, 375, 379, 383, 439, 447, 479, 495, 511] The trace representation of u_i is 7:

$$U(x) = Tr(\alpha^{462}x + \alpha^{858}x^3 + \alpha^{330}x^5 + \alpha^{363}x^7 + \alpha^{594}x^9 + \alpha^{528}x^{11} + \alpha^{66}x^{13} + \alpha^{429}x^{17} + \alpha^{693}x^{19} + \alpha^{264}x^{21} + \alpha^{858}x^{25} + \alpha^{363}x^{69} + \alpha^{528}x^{73}) \ . \quad (7)$$

The final trace form of the annihilator is mentioned as equation 8:

$$G(x) = Tr(\alpha^{66}x + \alpha^{858}x^3 + \alpha^{693}x^5 + \alpha^{429}x^9 + \alpha^{429}x^{17}), x \in F_{2^{10}} \ . \quad (8)$$

3) Online Mode Computation: The equation 8 will be used to recover the initial state or bits of the cipher. The bits captured by the adversary are now used to perform fast selective DFT attack. The initial state of the LFSR can be represented as:

$$w_0 = Tr(\beta), w_1 = Tr(\alpha\beta), w_2 = Tr(\alpha^2\beta), w_3 = Tr(\alpha^3\beta), w_4 = Tr(\alpha^4\beta), w_5 = Tr(\alpha^5\beta), w_6 = Tr(\alpha^6\beta), w_7 = Tr(\alpha^7\beta), w_8 = Tr(\alpha^8\beta), w_9 = Tr(\alpha^9\beta) \quad (9)$$

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

To find out the initial state, β needs to be computed and captured keystream bits are utilized for this purpose. All bits index i where bits are 1 will be used to form system of equation on F_2^{10}. The equation 8 is substituted for $x = \alpha^i\beta$ to form the equation 10:

$$Tr(\alpha^{66+i}\beta + \alpha^{858+3i}\beta^3 + \alpha^{693+5i}\beta^5 + \alpha^{429+9i}\beta^9 + \alpha^{429+17i}\beta^{17}) = 0\ . \quad (10)$$

All corresponding β^n is replaced with equation11 in equation 10, for all n representing cyclotomic cosets of F_{2^5} to form System of equation 12:

$$\beta^n = x_0 + \alpha x_1 + \alpha^2 x_2 + \alpha^3 x_3 + \alpha^4 x_4 + \alpha^5 x_5 + \alpha^6 x_6 + \alpha^7 x_7 + \alpha^8 x_8 + \alpha^9 x_9\ . \quad (11)$$

$$Tr(\alpha^{66+i}(x_0 + \alpha x_1 + \alpha^2 x_2 + \alpha^3 x_3 + \alpha^4 x_4 + \alpha^5 x_5 + \alpha^6 x_6 + \alpha^7 x_7 + \alpha^8 x_8 + \alpha^9 x_9)) + Tr(\alpha^{858+3i}(x_0 + \alpha x_1 + \alpha^2 x_2 + \alpha^3 x_3 + \alpha^4 x_4 + \alpha^5 x_5 + \alpha^6 x_6 + \alpha^7 x_7 + \alpha^8 x_8 + \alpha^9 x_9)) + Tr(\alpha^{693+5i}(x_0 + \alpha x_1 + \alpha^2 x_2 + \alpha^3 x_3 + \alpha^4 x_4 + \alpha^5 x_5 + \alpha^6 x_6 + \alpha^7 x_7 + \alpha^8 x_8 + \alpha^9 x_9)) + Tr(\alpha^{429+9i}(x_0 + \alpha x_1 + \alpha^2 x_2 + \alpha^3 x_3 + \alpha^4 x_4 + \alpha^5 x_5 + \alpha^6 x_6 + \alpha^7 x_7 + \alpha^8 x_8 + \alpha^9 x_9)) + Tr(\alpha^{429+17i}(x_0 + \alpha x_1 + \alpha^2 x_2 + \alpha^3 x_3 + \alpha^4 x_4 + \alpha^5 x_5 + \alpha^6 x_6 + \alpha^7 x_7 + \alpha^8 x_8 + \alpha^9 x_9)) = 0\ . \quad (12)$$

From the linearity of the trace function, system of equation is solved to extract values of $[x_0, \cdots x_9]$. Which is finally substituted in 11 to get value of β to recover the initial state.

TABLE I
COMPARISON OF KEY RECOVERY ATTACKS ON WG-7

Complexity	DFT Attack [10]	AA [11]	FAA [11]	Our Attack
Offline Computation	$O(2^{39.5})$	-	$O(2^{26.87})$	$O(2^{41})$
Keybits Requirements	$2^{25.5}$	$2^{19.38}$	$2^{19.3}$	2^{18}
Online Computation	O($2^{25.5}$)	O($2^{54.36}$)	O($2^{26.73}$)	O(2^{21})

D. Comparison of Key Recovery Attacks on WG-7 Complexities

Table I compares the key recovery attack results. Column 1 gives the predicted DFT attack by Yiyuan et. al. [10], Algebraic and FAA attack by Orumiehchiha et al. [11] and proposed fast selective DFT attack with annihilator. It is evident by comparison of the complexities that proposed fast selective DFT attack with annihilator is efficient in online phase and keybits requirement than Yiyuan et. al. [10] predicted DFT attack, however, offline phase is not efficient than the former. The proposed method is also efficient in online phase and keybits requirement than Algebraic and FAA attack.

V. CONCLUSION

This paper has investigated the security of the WG-7 Stream cipher by proposing the use of annihilator in Fast Selective DFT Attack. New results of Discrete Fourier Transform attack against the stream cipher WG-7 have been presented which is an improvement over the previous attack. As a result, the complexity of DFT attack on the cipher has reduced considerably. The key recovery attack employing Fast Selective DFT attack recovers the key with key streams requirements of about 2^{18} with online computation of about $O(2^{21})$. The presented results reveals that WG-7 stream cipher is not secure against DFT attack. The same attack can be extended to WG-8 [12] cipher after finding out the annihilator of non-linear filter function.

REFERENCES

[1] S. Rnjom and T. Helleseth, "A new attack on the filter generator," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1752–1758, 2007.

[2] ——, "Attacking the filter generator over gf(2^m)," in *Arithmetic of Finite Fields, First International Workshop, WAIFI 2007, Madrid, Spain, June 21-22, 2007, Proceedings*, Madrid, Spain, 2007, pp. 264–275.

[3] Y. Nawaz and G. Gong, "WG: A family of stream ciphers with designed randomness properties," *Inf. Sci.*, vol. 178, no. 7, pp. 1903–1916, 2008.

[4] G. G. Sondre Rnjom and T. Helleseth, "On attacks on filtering generators using linear subspace structures," in *Sequences, Subsequences, and Consequences, International Workshop, SSC 2007, Los Angeles, CA, USA, May 31 - June 2, 2007, Revised Invited Papers*, Los Angeles, CA, USA, 2007, pp. 204–217.

[5] G. Gong. (2011) A closer look at selective dft attacks. Department of Electrical and Computer Engineering; University of Waterloo. Waterloo, Ontario, Canada. [Online]. Available: http://cacr.uwaterloo.ca/techreports/2011/cacr2011-35.pdf

[6] T. H. Guang Gong, Sondre Rnjom and H. Hu, "Fast discrete fourier spectra attacks on stream ciphers," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5555–5565, 2011.

[7] K. C. Jingjing Wang, Xiangxue Li and W. Zhang, "Attack based on direct sum decomposition against the nonlinear filter generator," in *Progress in Cryptology - AFRICACRYPT 2012 - 5th International Conference on Cryptology in Africa, Ifrance, Morocco, July 10-12, 2012. Proceedings*, Ifrance, Morocco, 2012, pp. 53–66.

[8] J. Wang and K. Chen, "Improvement of discrete fourier transform attack," *Journal of shanghai Jiaotong University (Science)*, vol. 46, no. 2, pp. 285–288, 2012.

[9] K. C. Jingjing Wang and S. Zhu, "Annihilators of fast discrete fourier spectra attacks," in *Advances in Information and Computer Security - 7th International Workshop on Security, IWSEC 2012, Fukuoka, Japan, November 7-9, 2012. Proceedings*, Fukuoka, Japan, 2012, pp. 182–196.

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

[10] G. G. Yiyuan Luo, Qi Chai and X. Lai, "A lightweight stream cipher wg-7 for rfid encryption and authentication," in *Proceedings of the Global Communications Conference, 2010. GLOBECOM 2010, 6-10 December 2010, Miami, Florida, USA*, Miami, Florida, USA, 2010, pp. 1–6.

[11] J. P. Mohammad Ali Orumiehchiha and R. Steinfeld, "Cryptanalysis of WG-7: a lightweight stream cipher," *Cryptography and Communications*, vol. 4, no. 3-4, pp. 277–285, 2012.

[12] K. M. Xinxin Fan and G. Gong, "WG-8: A lightweight stream cipher for resource-constrained smart devices," *ICST Trans. Security Safety*, vol. 3, p. e4, 2015. [Online]. Available: http://dx.doi.org/10.4108/sesa.2.3.e4

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

A Comparison of GPU based Networking Algorithms

Jaideep Patel
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India

Nilay Khare
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India

***Abstract*— Modern high performance routers rely on customized hardware solutions, therefore difficult to adapt to continuously changing network protocols. While software routers provide the flexibility and programmability, thereby achieving a lower throughput. Modern GPUs offer significant computing power, and its data parallel computing matches the packet processing on routers. As the routing table size increases along with the physical link speeds, IP address lookup has been a difficult problem at routers. It has been a challenge to achieve a cost effective high-performance IP lookup. With the increasing Internet traffic and fast changing network protocols, the next generation routers have to satisfy the need for throughput, scalability, flexibility and QoS. The aim of this paper is to compare different GPU based networking algorithm on different performance metrics and to discuss their benefits and drawbacks.**

Keywords-Networking algorithms; Software Router;GPU; IPLookup; QoS.

I. INTRODUCTION

The internet and wireless technologies are showing rapid growth and have been immensely successful in recent years, but due to increase in network size and no. of users, network congestion is a recurring phenomenon which causes longer delay, packet loss and other performance related degradation. Computing is entering an era where an application needs lots of task performance. Such applications include multimedia applications, high frequency sensing applications, file transfer etc. Since the device utilizing these applications need to be an integral part of the network, solutions for reducing the effects of congestion in wireless networks are required. Accordingly Internet routers, have to deliver enhanced processing capacity. Modern routers need to perform data intensive applications such as intrusion detection along with tradition router applications such as packet classification and route table lookup which cause the challenge in router throughput. To move in synchronization with the growing Internet traffic, high performance routers depend on custom hardware. On the other hand, proprietary router hardware does not meet the requirement for programmability. The custom hardware solutions make it hard for routers to get used to varying network protocols .On modern routers, packet processing involves a series of operations on packet headers by a CPU or a network processor. A router has to deliver a high throughput under rigorous quality-of-service requirements. The existing protocols have to be updated to become accustomed to changing network applications.

Graphic processing units (GPUs) are emerging as a new trend in high-performance computing. Multi-core CPUs generally utilize the parallel processing at the task level. The GPU computing is an example of parallel data programming. A GPU would launch multiple threads at the same time for a particular work, each thread processing the same program but on a different set of data.

It is advised to use GPUs as a dedicated pocket processor in a software router. There are many advantages of using GPU in software router. Superior computing power of GPU can be utilized for high packet processing output. Also, the software router can assembled with off-the-shelf hardware making it cost efficient. Another advantage is that GPUs are supported by mature software development tools as these tools address a mass market. The remaining paper is organized as follows: Section 2 gives the description about the GPU based networking algorithms. Section 3 performs Comparison and Analysis. Section 4 provides the conclusion of the paper.

II. GPU BASED NETWORKING ALGORITHMS

A. IP Routing Processing with Graphic Processors

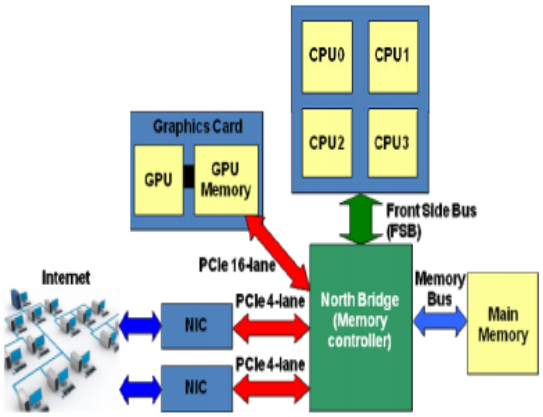


Figure 1. GPU based software Router

Mu [1] suggested GPU solutions for a chain of core IP routing applications like IP routing table and pattern match. Bloom Filter and Finite Automata based matching algorithm were used for deep packet Inspection.GPU based solution for routing table lookup is suggested in this work.Fig.1 and Fig.2 shows

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

the GPU based software router and work flow in processing matching respectively.

- Network Intrusion and detection

  In current network devices, network intrusion techniques are being used to cater to demand of services like shaping of traffic, service quality, and detection of intrusion. The challenge is to match signature, which confirm if network payloads contain the signatures shared beforehand at line rates. In signature matching predefined rule set of strings is searched in a packet in the context of DPI. Mu[1] developed capable GPU solutions for the string and normal expression matching. There are two algorithms which are used in this work

  1) Bloom filter

  A Bloom filter is a simplification of hash table. It gives a simple, space-efficient data structure to represent a data set for fast membership queries. It is a data structure designed to ascertain, rapidly and memory-efficiently whether an element is present in a set. The Bloom Filter algorithm is very suitable for a GPU implementation since the evaluations on different strings are totally independent. GPU based Bloom filter processing has a throughput of about 19Gbit/s, 30 times faster than throughput of CPU (0.6Gbit/s).

  2) Aho-Corisick (AC)

  The AC algorithm utilizes Deterministic Finite Automata (DFA) for identifying regular expressions patterns. A DFA can be expressed with the help of a graph. In this graph each nodal point represent a unique state and an edge denotes a probable transition between states. An efficient data structure for the Deterministic Finite Automata is a 2-D transition table; where row represents to a state and column correspond to a dissimilar letter in acceptable alphabet.For the AC implementation, a throughput of 3.2 Gbits/s is achieved in paged-lock GPU kernel which is almost 5 times faster than CPU (0.6Gbit/s).

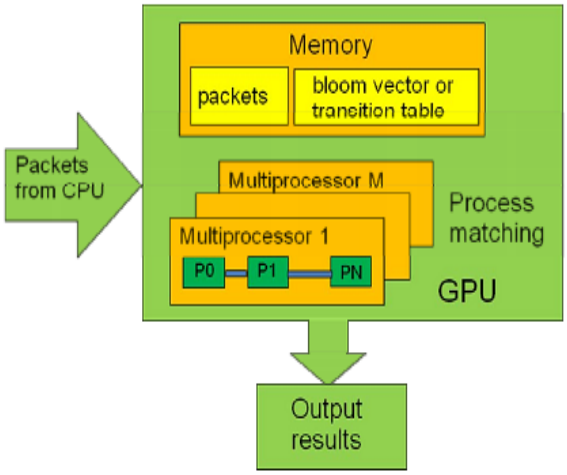


Figure 2. Working Flow in Processing Matching

- Routing Table Lookup for packet forwarding

  Longest Prefix Match algorithm has been utilized and extended further for GPU implementation. The routing table is structured like a network structure in which a node correspond a state in the search process and line connecting nodes represent bits value in the destination IP address. The movement or navigation across the network path represent and matching exercise. The Route table is taken care by the CPU. Whenever the route table is created or changed, it will be moved to GPU memory. The parallel organization is in that case inconsequential. For processing the IP in one packet one thread was used. Pointers handle nodes and edges in the radix tree. This kind of pointer chasing process is very hard in GPUs. A major problem is that the within the tree structure, pointers cannot be straightly moved to GPU memory. In the implementation, they suggested a new modified data structure called as “portable routing table” (PRT). PRT uses displacement in place of pointers for tree operations. The routing table is kept in the texture memory to utilize GPU’s on-chip cache. Proposed GPU implementation achieved a 6 times faster results. This paper developed efficient GPU implementations for the router applications that include IP routing table lookup and matching of pattern for network intrusion detection. Results displayed that GPU can accelerate packet processing by one order of magnitude.

B. PacketShader: A GPU-Accelerated Software Router

PacketShader utilizes the extensive-parallel processing power of GPU to resolve the CPU bottleneck in current software routers. Utilizing the high-performance packet I/O engine, PacketShader performs better than the existing software routers by a factor of four. Han[2] have done IPv4 and IPv6 forwarding, OpenFlow switching, and IPsec tunneling to exhibit the performance improvement of PacketShader.

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

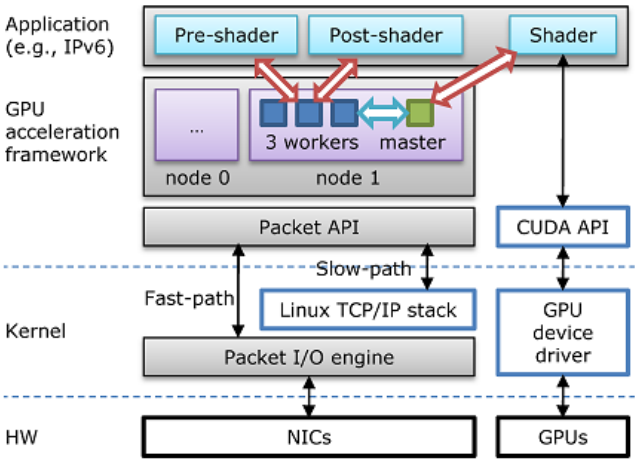


Figure 3. PacketShader software architecture

Figure 3 describes the architecture of PacketShader. In proposed work first, they implement highly optimized packet I/O engine which eliminate memory management overhead for each packet and for batch processing of enable high-performance packet I/O in user mode. In this process they use compact metadata and huge packet buffer, which reduces the initialization cost of the metadata structure and eliminates the cost of per-packet buffer allocation respectively.

Processing overheard per-packet is reduced significantly by performing batch processing of multiple packets. They extended batch processing to the application (packet processing) level .To maximize benefit from huge packet buffers they perform aggressive software prefetch. The compulsory cache miss latency is eliminated by this prefetch of the consecutive packets. Multi-core scalability and NUMA scalability of packet I/O has also described.

Second, they offload core operations (ex. IP table lookup or IPsec encryption) to GPUs and scale packet processing with immense parallel processing. Combined with I/O path optimization, GPU utilization is maximized by PacketShader by using as large parallelism as possible in a small time frame. PacketShader is a multithreaded program running in user mode. For Packet I/O, PacketShader invoke the packet API which consists of wrapper functions to kernel level packet I/O engine. A packet processing application runs on a framework and is driven by three functions (pre-shader, shader, post-shader)

Pre-shading: Each thread obtain a group of packets from its RX queues. It drops any incorrect packets and classifies normal packets that will be processed with GPU. Then it passes the input data to input queue of a master thread.

Shading: The master thread move the input data from host memory to GPU memory, execute the GPU kernel and returns results from GPU memory to host memory. Then it puts the results back to the output queue of the worker thread for post shading.

Post-shading: A worker threads takes up the results in its output queue and perform modification on drops or duplicates packets in the chunk depending on the result of the processing. Finally it splits the packets in the chunk into destination ports for packet transmission.

PacketShader uses optimization possibilities in the workflow.

- Chunk Pipelining
- Gather/Scatter
- Concurrent Copy and Execution

They execute four applications on top of PacketShader, which are IPv4 and IPv6 forwarding, OpenFlow switch and IPsec tunneling. Each application is executed in two modes, CPU mode and the CPU+GPU mode, which evaluate the effectiveness of GPU acceleration. PacketShader forwards IPv4 packets at pace of 40 Gbps for all the size of packets. GPUs brought major performance enhancement for memory and compute-intensive applications.

The per-packet processing overhead is minimized in network stack and packet processing without severe penalty is obtained. PacketShader can be highly scalable with multi-core CPUs, high-speed NICs, and GPUs in NUMA systems. The effectiveness of the approach is exhibited with IPv4 and IPv6forwarding, OpenFlow switching, and IPsec tunneling. The evaluation results displays that GPU gives significantly higher throughput over the CPU, assuring the usefulness of GPU for computation and memory-intensive operations in packet processing In this work they have demonstrated that PC-based router can achieve 40 Gbps forwarding performance with full programmability on commodity hardware.

C. Achieving O (1) IP Lookup on GPU-based Software Routers

IP address lookup poses a challenge because of the increasing routing table size, and higher line rate. Zhao [3] proposed novel ways to develop proficient IP lookup method using GPU. Their contribution lies in designing a fundamental architecture for high-performance IP lookup engine with GPU, and in developing efficient algorithms for routing prefix operations like lookup, deletion, insertion, and modification. The IP lookup scheme can attain O (1) time complexity.

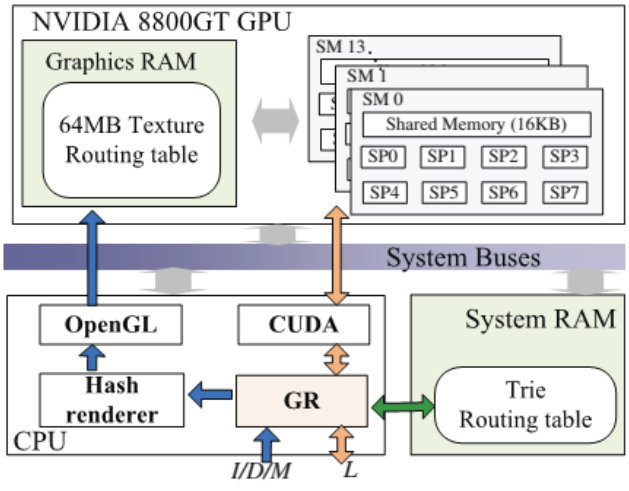


Figure 4. GPU-based IP lookup engine

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

Their GPU-accelerated IP lookup architecture GR is shown in Fig.4.

IPLookup: They proposed a square hash mapping policy in routing lookup. In this hash mapping policy the hash function takes a 24-bit number input, and provides 2D coordinate as output. In this hash function every prefix entry uses a square area in the texture, which can be proficiently manipulated by graphics processing facilities.

Modification: A lookup operation will be performed first to modify the next-hop information of a given entry. GR will write the new next-node information to the target memory after locating the memory address.

Insertion: The process of inserting a prefix entry is simple:

(1) put the GL depth function to GL GEQUAL, (2) Generate prefix's square (3)Draw the square by depth=prefix's slash value.

Deletion: Replacing the value of deleted entry with its parent's value which is performed in two passes: In the first pass, the deleted square's stencil buffer is set to zero, In the second pass the deleted square's stencil's buffer is made equal to its parent's depth value.

In this paper, they provided an alternative approach for building cost-effective high-performance IP lookup engines and the appropriate routing table update schemes using GPUs. Along with the parallelism, IP lookup method that could provide O (1) time complexity for each IP lookup. The key improvement contain: (1) IP addresses translated in memory addresses using a huge routing table on GPU memory. (2)They also perform the route update operations by making use of GPU's graphics processing facilities like Z-buffer, Stencil buffer.

They presented the design and evaluation of GR that exploits the immense parallelism to speedup routing table lookup and exhibit that there is the potential for significant improvement, e.g., 6 times faster speed than trie-based implementation. They also designed better algorithms that can perform deletion, insertion and modification operations.

D. Exploiting Graphics Processors forHigh-performance IP Lookup in Software Routers

IP lookup is a demanding problem at routers due to increase in the physical link rate and the routing table size. Demands for cost effective high-performance IP lookup has been growing. Conventional approaches normally use specialized hardware, such as TCAM. While these approaches are advantageous in terms of hardware parallelism in achieving high-performance lookup, they incur a high cost. Zhao et.al. [4] gives the contribution which lies in designing architecture for high-performance lookup engine with GPU, and in developing efficient algorithms for routing prefix update operations. Leveraging GPU's parallelism, the proposed schemes addressed the challenges in designing IP lookup at GPU-based software routers.

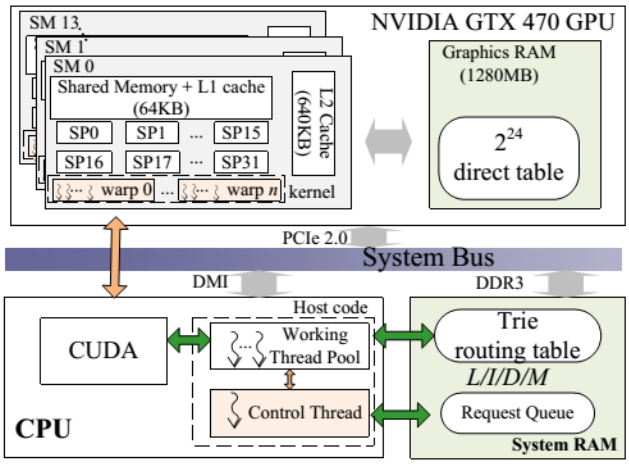


Figure 5. GPU-based IP lookup engine

The proposed GPU-accelerated IP lookup engine, GALE, as shown in Fig. 5

Zhao [4] used a new method for building cost-effective and high speed IP lookup engines and the corresponding routing table update methods using GPUs. Building an efficient IP lookup engine on GPUs, however, is a non-trivial task because of the difficult data-parallel programming model provided by the GPU, and also due to the significant demand in performance. GALE, a GPU-Accelerated Lookup Engine, allows high-performance IP lookup inside software routers. They made use of the CUDA [23] to facilitate IP lookup on GPU in parallel manner. They also utilize an IP lookup method that could achieve O (1) time complexity for IP lookup. The key idea include: (1) IP addresses are converted into memory addresses with the use of a large direct table on GPU memory which stores all possible routing prefixes. To find the next-node information only one direct memory access is requisite. This scheme can therefore be scaled to extremely high line rate with tiny computing overhead and latency for accessing the memory. (2)The large routing table although efficient for lookup, makes routing updates, which happens regularly in core routers, much more time-consuming and complex. Multiple prefixes need to be updated when only original prefix is added or deleted. This can be resolved by performing routing update operations on GPU.

Lookup:

Proposed work try to place the next node value of every likely IP prefixes in the direct table. In this way each IP address has one-to-one mapping with entries in direct table. Direct table‘s entries are derived from trie structure. In a lookup for the incoming IP address for example a.b.c.d the left most significant 24 bits a.b.c are used as index to its corresponding next hop information in the direct table.

Routing table update:

a)Insertion and Modification:

Insert and modify operation are basically the same operation for direct table. In a new route prefix, GALE will update the new next node value to the appropriate IP prefixes. The update

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

ranges rely on length of new prefix. However, the update for the next node values rely on condition of new prefix length is more than the existing prefix length.

b)Deletion:

The delete operation is similar to modify operation except for the period of update operation of the range of entries, the next node information is modified to the parent node's nexthop information in the trie. Therefore, deleting an entry is done as follows: replacing the next node and prefix length of the updated entry with corresponding value of parent's next node and prefix length. The parent node is attained by the trie-traversal for the period of eliminate the entry node in trie by CPU.

Designing efficient IP lookup algorithms in software routers that can meet the fulfill objectives of cost-effectiveness, high-performance lookup and low cost route-update is a challenge. To address above challenge, the following key improvement has been made in this paper: (1) they proposed the design and evaluation of a GPU-accelerated IP lookup engine, GALE, which utilize the immense parallelism for speeding up the parallel routing table lookups. (2) Authors also propose using a direct table for effective IP lookup with small computing overhead and memory access latency. Specifically, it can obtain O(1) computational complexity and only one memory access for each lookup request. (3)They finally designed efficient methods that perform route-update operations like deletion and insertion. Experiment results demonstrated that, there is the potential for significant improvement in lookup throughput using a proper design. With a better performance/price ratio, and very much flexible programmability, GPUs are ready to showcase high-performance routing processing.

E. Hermes:An Integrated CPU/GPUMicroarchitecture for IP Routing

With the ever growing Internet traffic and rapid varying network protocols, future routers have to fulfill the demand for throughput, scalability, QoS, and flexibility. Zhu [5] proposed a new integrated CPU/GPU micro architecture, Hermes, for QoS-aware high speed routing. Zhu [5] in addition proposed a new thread scheduling mechanism, which improves QoS metrics significantly.

1) Hermes is an incorporated CPU/GPU shared memory micro architecture contains an adaptive warp issuing mechanism for IP packet processing. For optimizing the QoS metrics Zhu [5] developed a GPU based packet processing platform.

2) This architecture is used to implement router related application. Authors did extensive performance evaluations of QoS. Hermes gives 5X improved throughput compared to GPU based router [1].Also an 81.2% lessening in average packet delay, as well as a 72.9% decrement in delay variance.

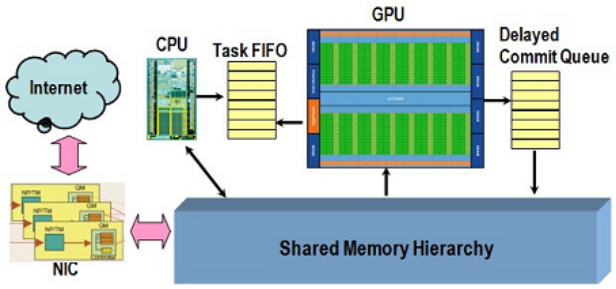


Figure 6. Hermes Micro architecture

A diagram of Hermes is shown in Figure 6.The basic execution flow remains the similar to a CPU/GPU system. The packet data are stored into the shared memory on arrival and then retrieved by shader cores for processing. The processed packets are updated in the shared memory, from where they can be further processed by CPU or forwarded.

Adaptive Warp Issuing Mechanism:

The warp issuing mechanism of Hermes handles assignment of parallel tasks to shader cores for any further intra-core scheduling. Currently all the thread warps in GPU are kept in a warp pool before being issued.

For packet processing applications, it is unavoidable to wait for an enough number of warps under certain circumstances. Therefore, they proposed a mechanism which adapts to the pattern of arrival of network packets and maintains equilibrium between overall throughput and worst-case per-packet delay.

In the fine-grained multithreading execution model, thread warps running on one shader or different shader core may finish in an arbitrary order. As a result, processing can happen in any order for sequentially arrived packets. Protocols like UDP need in-order processing so they use Delay Commit Queue (DCQ).

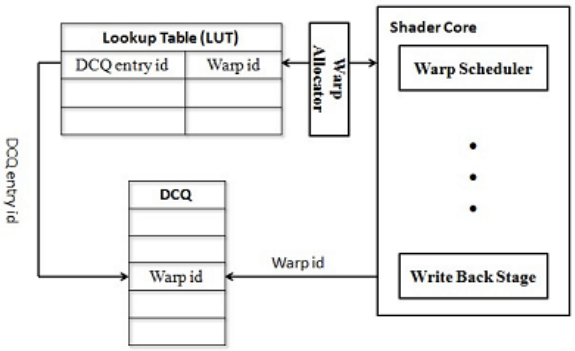


Figure 7. Delay Commit Queue with a Lookup Table

The key idea is to allow out-of-order warps execution but enforce in-order commitment. In Figure 7, Delay Commit Queue holds the IDs of the warps which have finished although not committed yet. Each time the warp is going to issue on core, and if DCQ have the space, new entry will be assigned. Mapping between warp ID and the DCQ entry ID is placed in a LUT. When finishing, the appropriate DCQ entry is modified

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

by indexing the Lookup table, with the completed warp's ID. When all warps arrived previously has been finished only then a warp could be committed. When a warp commits, its DCQ entry will be reclaimed.

A complete set of router applications were implemented on the Hermes architecture. Experimental results showed that this architecture could meet strict delay requirements, maintaining a high processing throughput at the same time.

F. Scalable Packet Classification via GPU Metaprogramming

Packet classification is a primary processing pattern of modern networking devices. contemporary routers use special hardware meant for packet classification, however such solutions are not cost efficient, have high power consumption and poor extensibility, while software-based routers gives the best flexibility, but deliver less performance . GPUs have proved to be an effective accelerator for software routers. Kang[6] proposed a GPU-based linear search framework for packet classification. The crux of this framework is a programming technique that significantly enhances the execution efficiency.

GPU packet classification using DBS algorithm:

DBS algorithm was implemented on both GPU and CPU. Preparation was conducted on CPU in both versions. GPU threads are organized into 120 blocks, each consisting of 128 threads. 14,336 packets are processed during each kernel launch in a single batch.

In DBS the partitioning of rule sets is limited to certain granularities, which suggest a performance bound for more general algorithm categories than only DBS or hash-based algorithms. The limit is due to the overlapping of rules in the classification space.

GPU packet classification using linear search:

1) Each rule is translated as a programming logic in C which checks if the packet matches. The rules are now embedded as compilation-time constants.

2) The series of kernels is converted to code fragments, which is compiled and upload to the graphics card in binaries.

3) Packet header data is loaded into GPU memory, kernels are executed, and then classification results are moved back to CPU.

This technique eliminate fetching of rules from memory; the rule set is compiled as a C program code, which can be effectively broadcasted to all CUDA cores on a GPU with the instruction issuing approach. The memory latency for instruction binary can be hidden by GPU's instruction cache.

Authors investigate effective data-parallel algorithms for packet classification, following the direction of previous works on GPU-based software routers [1] [2]. A main observation is that many algorithms have low theoretical complexity but have poor scalability with large rule set and does not meet the performance constraints .Authors resolved the scalability issue by focusing on efficient parallelization and avoiding complicated algorithm. They proposed a GPU-based linear search scheme for packet classification by adopting a novel meta programming approach. Experimental results demonstrate that it significantly outperforms existing software solutions, even for large rule sets.

GPU linear search gives 17 X higher throughput than Discrete Bit Selection algorithm on CPU or 8.5 X higher than its own CPU counterpart. The prototype implementation offers significantly higher classification throughput than its CPU counterpart, while at the same time maintains good scalability even when the size of rule set reaches 50,000. They also demonstrated that meta programming can be a powerful tool for GPU computing.

G. An Analysis of Queuing Network Simulation Using GPU-Based Hardware Acceleration

Queuing networks finds its application broadly in computer simulation studies. Queuing networks are supply chains, manufacturing and routing. If the networks are small in size and complexity, discrete event simulations of the networks can be created without incurring major delays in analyzing the system. However, as the networks size grows, such kind of analysis can be time consuming exercise and thus need expensive parallel processing computers or clusters.

Park et.al.[7] have developed tools that simulate queuing networks in parallel, using the fairly economical and generally available graphics processing units (GPUs) found in most recent computing platforms. They gave an analysis of a GPU-based algorithm, highlighting pro and cons with the GPU approach. An algorithm cluster events, attaining speedup at the expense of an estimate error which grows as size of cluster enhance. They were able to achieve 10-x speedup using this approach with only a small error in a particular implementation of a closed queuing network simulation. Error can be reduced, base on error analysis tendency, thus achieving reasonably correct output statistics. The experimental results of the MANET simulation displays that errors happen merely in the time-dependent output statistics.

Authors examined three different queuing models to study the effects of their simulation method: the general network topologies of open and closed queuing networks, and computer network models. They analyzed the trade-offs between numerical errors and performance gain and the methods for error estimation and correction. They used parallel execution for reducing the overall execution times in the simulation on the GPU.

Time Synchronous /Event Algorithm:

This algorithm is a mixture of discrete event simulations and time stepped simulations. In this approach they execute the events at the end of the time interval to increase the extent of parallelism. If events are executed only at the end of the time interval, each event will have delayed execution with respect to its original timestamp and the results will lose accuracy. They rely on queuing theory for estimating the total error rate of simulation results. A non decreasing timestamp order of execution of events is preserved by the synchronous step of

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

simulation, which blocks the event extractions from the FEL before the current events finish scheduling the next events.

Closed and open queuing network:

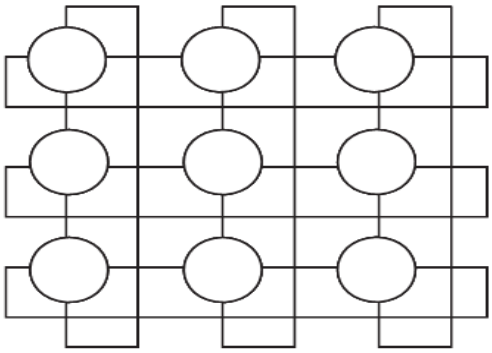


Figure.8 Closed queuing network

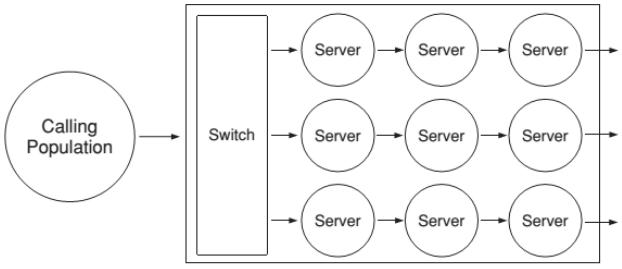


Figure 9. Open queuing network

Fig.8 and Fig.9 shows closed and open queuing networks respectively. In an Open queuing network the number of tokens in the network at any point is always different due to the arrivals and departures where as in closed queuing network the number of token are constant during the simulation.It is possible to have a fixed size array for the FEL in the closed queuing network because of the fixed no. of tokens during the simulation. In an open queuing network, the size of the FEL is always different, so size of FEL is set to double the size of tokens, due to the variable number of tokens at any instant of time.

Computer Network Model:

Queuing model were developed to analyze computer network systems performance. Fig.10 shows the Queuing delay in computer network model. There are four types of delays when a packet is sent from one node to adjoining node which are: processing, queuing, transmission and propagation delays. Queuing delay is most studied among all these delays as it is the only delay caused by the traffic load and congestion pattern.

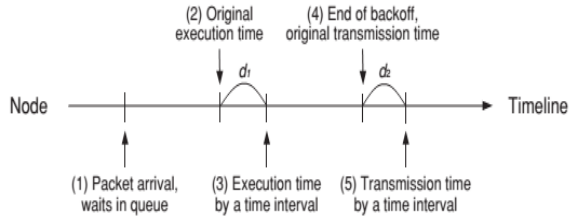


Figure 10. Queuing delay

Error Analysis:

They have found error to be frequent enough for approximation. So they developed error estimation and correction methods to provide more precise statistics. The total error rate equation is derived from the queuing theory equations. These equation coupled with the results of the simulation provide more accurate results without developing a complicated analytical model from each node. They analyzed their method impact on different queuing models. The problem of event distribution and inapt application of GPUs for discrete event simulation are dealt (1) by allowing events to occur at approximate boundaries at the cost of accuracy and (2) by using a detection and compensation method for error reduction. The experimental results exhibit 10 x speed up using their algorithm at the cost of accuracy in the results. The statistical results in MANET simulations show that their method only has an error in the time-dependent statistics. Although the improvement in performance caused error into the results of simulation, the experimental results showed that the error in the queuing network is frequent enough to estimate more accurate results, which can be reduced from queuing theory.

III. Comparison and Analysis

From the above GPU based networking algorithms it can be seen that they differ in terms of their application along with throughput, delay, QoS etc. As illustrated in Table 1, Different networking algorithms are executed in parallel manner with improvement so that achieved results of these algorithms are better than the existed ones.

IV. Conclusion and Future Work

Among several network devices, Internet routers play an important role by serving as the backbone. A router delivers packets between connected networks in a timely manner. Conventional routers rely on custom hardware for achieving high processing speed. Also being costly, such solutions are not adaptable to fast changing network services and diversified usage cases. On the other hand, software routers have been attaining attention of the researchers due to their extensibility and customizability. Software routers offer flexibility and cost-efficiency as they are built using commodity hardware. However, despite having better flexibility, software routers cannot meet the performance level required for growing Internet traffic. The application of software routers is limited because of this low performance. There has been a move in high-performance computing to carry out general purpose computing with graphic processing units (GPUs) due to its

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

cost efficiency and availability. For compute-intensive applications, the large number of array of GPU cores offers higher computation power. In this paper we compared different GPU based networking algorithms on different parameters. We have found that GPU has shown a substantial performance boost in networking algorithm and it will be the helpful for extending the original networking protocols to next generation high speed systems.

REFERENCES

[1] S. Mu, X. Zhang, N. Zhang, J. Lu, Y. S. Deng, andS. Zhang. , " Ip routing processing with graphicprocessors, " In DATE '10, March 2010.

[2] Sangjin Han,Keon Jang,KyoungSoo Park,Sue Moon "PacketShader: A GPU-Accelerated Software Router," SIGCOMM 2010.

[3] Jin Zhao, Xinya Zhang, Xin Wang, and XiangyangXue, "Achieving O(1) IP Lookup on GPU-based Software Routers,"SIGCOMM'10.

[4] Jin Zhao, Xinya Zhang, Xin Wang ,Yangdong DengXiaoming Fu, "Exploiting Graphics Processors for High-performance IP Lookup in Software Routers," 2011.

[5] Yuhao Zhu,Yangdong Deng, Yubei Chen., "Hermes: An Integrated CPU/GPU Microarchitecture for IPRouting,"DAC'11

[6] Kang Kang,Yangdong Steve Deng , "Scalable Packet Classification via GPU Metaprogramming ,"EDAA 2011.

[7] hyungwook park and paul a. fishwick, "An Analysis of Queuing Network Simulation Using GPU-Based Hardware Acceleration,"ACM 2011.

[8] AYANI, R. et.al.,"Parallel discrete event simulation on SIMD computers," Journal of Parallel. Distribituted computing 1993.

[9] P. Saif, P.W. Anderson, A. Degangi, "Gigabit routing on a software-exposed tiled-microprocessor," Architecture for networking and communications systems. 2005.

[10] C. Kulkarni, M. Gries, C. Sauer and K. Keutzer , "Programming challenges in network processor deployment ," Architecture and Synthesis for Embedded Systems Nov. 2003.

[11] E. Kohler, et.al. , " The Clock Modular Router," ACM 2000.

[12] P. Handley,et.al.,"Designing extensible IP router software ," Symposium on Networked Systems Design & Implementation. May. 2005.

[13] R. Bolla et. al., "Performance and power consumption modeling for green COTS Software Router," Communication Systems and Networks and Workshops,2009.

[14] J. Trank, "Packet Processing with Intel Multi-core Processors," Solution Brief, 2008.

[15] fujimoto et.al., "Exploiting temporal uncertainty in parallel and distributed simulations," IEEE 1999

[16] R. Simth et.al.,"Evaluating GPUs for network packet signature matching," Performance Analysis of Systems and Software 2009.

[17] G. Vasiliasdis, S. Antonatos, et al.,"Gnort: High Performance Network Intrusion Detection Using Graphics Processors," symposium on Recent Advances in Intrusion Detection 2008.

[18] Y. Luo,et.al., "Shared Memory Multiprocessor Architectures for Softwre IP Routers," IEEE. 2003

[19] M. Shin, Y. Kim, "New Challenges on Future Network and Standardization," Advanced Communication Technology Feb. 2008.

[20] K. Argyraki et. al., " Can software routers scale?," ACM 2008

[21] A. Bianco, R. Birke, "Control and management plane in a multi-stage software router architecture," High Performance Switching and Routing, May. 2008

[22] J. Owens et. al., "A Survey of General-Purpose Computation on Graphics Hardware," Eurographics (The European Association for Computer Graphics) and Blackwell Publishing. 2005

[23] http://www.nvidia.com

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

TABLE 1: Comparison of GPU based networking algorithms

Protocols	**Algorithms used**	**Performance metrics**	**Speed up**
IP Routing Processing with Graphic Processors[1]	(Bloom filter, Aho-Corisick), Routing Table Lookup	Throughput, Lookup time	30,5,6
PacketShader[2]	IPv4 Forwarding	Throughput	4
Achieving O(1) IP Lookup on GPU-based Software Routers[3]	IPLookup	Lookup time	6
GALE[4]	IPLookup	Lookup per second	5-6
Hermes:An Integrated CPU/GPU Microarchitecture for IP Routing[5]	IP Routing Processing, QoS evaluation	Throughput, delay , delay variance	5,81%,73%
Scalable Packet Classification via GPU Metaprogramming[6]	Linear search	Throughput with scalability	8.5
An Analysis of Queuing Network Simulation Using GPU-Based Hardware Acceleration[7]	Queuing Network Simulation	Event execution time	10

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

A Survey-Defect Detection and Classification for Fabric Texture Defects in Textile Industry

Ms. Shweta Loonkar
Ph.D Research Scholar,
Computer Engineering Department,
MPSTME, NMIMS University,
Mumbai-56

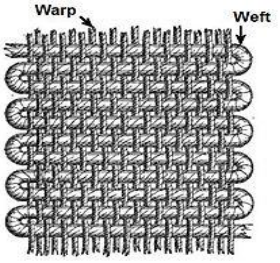
Dr. Dhirendra Mishra
Associate Professor,
Computer Engineering Department,
MPSTME, NMIMS University,
Mumbai-56

Abstract: Textile industry is one of the largest and oldest sectors in the India and has a formidable presence in national economy in terms of output, investment and employment. Due to increasing demand for quality fabrics it is thus important to produce the defect free high quality fabric. Visual inspection system consumes a lot of time and are error prone. The price of the fabric is reduced to 45%-65% due to presence of various defects. The purpose of this paper is to automate the detection and classification of texture defects by computerize software. The proposed method uses a statistical based approach for the inspection and detection of the defect on woven/knitted fabric collected from the textile industry. In this the images are acquired, pre-processed, restored and normalized to extract the statistical feature using computer vision. The extracted features are given an input to the artificial neural network decision tree classifier to compute the weighted factor for detecting and classifying the type of defects. An automatic defect detection system can increase the texture defect detection percentage and will reduce the fabrication and labour cost and improves the quality of the product.

Keywords: Defect detection, Statistical approach, Computer vision, Decision tree classifier, neural network.

I. INTRODUCTION

The Indian textile industry or apparel industry is one of the most important industry for the Indian economy. It primarily concerned with the production of yarn and cloth. Its importance is underlined by the fact that it accounts for around 4% of GDP, 14% of the industrial production and 17% of the country's total export earnings. All textile industries aim to produce competitive fabrics which is based on the productivity and quality of the fabrics produced by each industry. In the textile sector, there have been an enlarge amount of losses due to defective fabrics. A fabric defect is any abnormality in the fabric that hinders its acceptability by the consumer. Various types of faults present on knitted and woven fabrics are as hole, scratch, stretch, dirty spot, fly yarn, cracked point, slub (thickness of yarn), colour bleeding(poor wet fastness), dye spot, broken lines, knots, pin marks, missing yarn, thin yarn, thick yarn, and bad selvage etc. if these faults are not detected properly then they can affect the production process massively.



The fabric texture is usually made up of the repetition arrangement of warp and weft. The longitudinal threads are called the warp and the lateral threads are the weft or filling. The method in which these threads are inter woven affects the characteristics of the cloth and if not woven properly produces the defect in the fabric. A wide variety of defects are represented many defects are a direct cause of machine malfunction while others are from faulty yarns [9]. The various types of defects detected during quality controls are broadly classified as follows [16] [17]:

- Critical Defects – Defects which cause hazard to the health of individuals using it.
- Major Defects – More serious defects which are likely to affect the purchase of the product.
- Minor Defects – Include small faults which have no effects on the purchase of the product

Some of the commonly occurring fabric defects are discussed in the below mention table (I) from [4].

TABLE I. VARIOUS TYPE OF FABRIC DEFECTS

Yarn Defects	Woven Defects	Knitted Defects	Dyeing & Finished Defects
The defects originating from the spinning	The defects which originate during the process of weaving	The defects which occurs during knitting of cloth	The defect which occurs during the dyeing of textile products
Broken Filaments	Broken Ends	Drop Stitches	Shade Variation
Knots	Float	Yarn Streaks	Crease Mark
Slub	Gout	Barriness	Pin Hole Damage
Fabric press off	Hole, Cut or Tear	Fabric press off	Dye Spots
Broken ends	Oil Stain	Broken Ends	Wrong Slitting
Thick places	Slub	Spirality	Band Line
Thin places	Missing end	Slub	Dust
	Missing Picks	Pin Hole	
	Reed Mark	Broken Needle	
	Colour Bleeding	Cracks or Holes	

Manual defect detection in a fabric quality control system is a difficult task to be performed by inspectors. It has been observed [2] that price of textile fabric is reduced by 50% to 70% due to defects. The work of an observer is very tedious and time consuming. They have to detect small details that can be located in a wide area that is moving through their visual field. The identification rate is only about 70% [3]. So, early and accurate detection of defects in fabrics is an important aspect for product and quality improvement. Human visual inspection and automated inspection are

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

compared in Table II from [4]. Computer vision technology with artificial neural network have been applying on textured samples over the past few years for developing automated defect detection and classification systems. Automatic fabric defect detection systems mainly faces two challenging problems;

1. Defect Detection
2. Defect Classification.

Feature selection and image enhancement plays an important role in developing automated defect classification system. Image processing techniques deal with image acquisition, segmentation, manipulation, and analysis of images. The advantages of digital imaging are accurate data acquisition, better combination of spatial and contrast resolution, compact storage/easy retrieval, fast accurate image transmission. For selecting an appropriate feature set, the distinguishing qualities of the features should be high and the number of features should be small and takes into account the difficulties that lie in the feature extraction process [1]. Neural networks and decision tree classifiers [35] are suitable enough for developing real-time systems because of their parallel-processing capability. Moreover, NNs have strong capability and good accuracy to handle multiclass classification problems. Classification accuracy, model complexity and training time are three of the most important performance metrics of NN models.

TABLE II. COMPARISON OF HUMAN AND AUTOMATED DEFECT INSPECTION

Inspection Type	Visual Inspection Versus Automated Inspection	
	Visual	**Automated**
Fabric Types	100%	70%
Defect Detection	70%	80%+
Reproducibility	50%	90%+
Objective Defect Judgement	50%	100%
Statistics Ability	0%	95%+
Inspection Speed	30m/min	120m/min
Response Type	50%	80%
Information Content	50%	90%+
Information Exchange	20%	90%+

The remaining section of this paper is organized as follows. Section 2 describes relevant previous efforts in the fields, such as textile fabric inspection systems, computer vision and machine learning systems like neural networks for automated textile defects recognizing etc. Section 3 provides the research objective and approach used to automate the proposed textile defect detectors. Section 4 describe the proposed system in detail including the method for defect extraction, the scheme of feature extraction, and the structure of the neural network decision tree classifier. Section 5 gives the discussion and observations. Section 6 concludes with some remarks and plausible future research lines.

II. LITERATURE REVIEW

Fabric defect detection using digital inspection images has received considerable attention during the past decade and numerous approaches have been proposed in the literature [1-45]. At microscopic level, the inspection problems encountered in digital images become texture analysis problems [5]. Revathy-Vijaylakshmi [6] obtained the principal components from the co-occurrence matrix of fabric texture and further fuzzy rule based classification is done. It has been found that by using thresholding techniques 90% of the defects in a plain fabric could be detected and have used resilient backpropagation algorithm to train their NN [7][8]. Their networks have been capable of dealing with multiclass problem. Sebeanni [9] and Ajay [5] have introduced gray-level statistical method for detecting the defects from the textile fabrics. Mac-Yiu [10] and [11] has given morphological filters techniques to tackle the problem of automated defect detection for woven fabrics. In the proposed scheme, important texture features of the textile fabric are extracted using a pre-trained Gabor wavelet network [10]. These methods depend on intensity change on the fabric image, can only capture significant defects such as knot, web, and slub. The reduction of wastage, higher price of fabrics due to the presence of fewer defects, requirement of less labour, and other benefits make the investment in an automated textile defect inspection system economically very attractive. The development of a fully automated web inspection system requires robust and efficient defect detection and classification algorithms [28] [29]. The inspection of real textile defects is particularly challenging due to the large number of textile defect classes, which are characterized by their vagueness and ambiguity [12]. The justification for fabric defects could be described to the fact that no production or manufacturing process is 100% defect-free which applies particularly where natural materials, as textile ones, are processed [13]. There are several reported works [14, 15, 16, 17, 18 and 20] discuss the influence of fabric defects on textile industry.

In the last two decades, there have been several key developments in automated visual inspection technique for fabric defects where new approaches such as an ultrasonic imaging system [19] have also been proposed.

To deal with challenges of machine vision based fabric inspection system, numerous attempts have been made all around the globe in developing techniques to detect and classify fabric defects [3–20]. Most of them have concentrated on defect detection, where few of them have concentrated on defect classification. There have been deployment of mainly three defect-detection techniques [12], namely, statistical, spectral, and model-based. A number of techniques have been deployed for classification. Among them, neural network, support vector machine (SVM), clustering, and statistical inference are notable. But, the main common alternative to human visual defect detection is the use of a computer vision system to detect differences between images acquired [20]. This means that texture analysis plays an important role in automatic visual inspection of surface [17, 21, 22, and 23] features.

It is, however, worthwhile to recall that fabric defects are loosely separated into two types [17]; one is global deviation of colour (shade); the other is local textural irregularities which is the main concern for our study. Co-occurrence

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

matrix technique is based on different grey level configurations in a texture fabrics [16, 24]. This co-occurrence technique can be computationally expensive for the demands of a real-time defect inspection system. The number of gray levels is usually reduced in order to keep the size of the co-occurrence matrix manageable [34]. The process at which these defects are detected is called fabric inspection. In addition, the visual inspection has worked well for many years in part because the amount of data has been small and manageable. Automatic inspection systems are designed to increase the accuracy, consistency and speed of defect detection in fabric manufacturing process to reduce labour costs, improve product quality and increase manufacturing efficiency [1-37]. We will categorise the texture analysis problem into four approaches according to the used algorithm; structural, statistical, spectral, and model-based approaches. Among all statistical approaches are very popular. The above part of the literature presents in Table III summarizes our modified automated texture defect detection and classification techniques and the various algorithms used in the research.

TABLE III. LITERATURE REVIEW OF AUTOMATED TEXTURE DEFECT DETECTION AND CLASSIFICATION

Approaches	Methods	References
Structural Approaches		[25,17,16]
Statistical Approaches	Grey Level Thresholding	[16,17,26, 24,18]
	Cross Co-relation	[16, 17, 40]
	Statistical Moment	[17, 22, 23]
	Histogram Equalization	[25, 22, 6]
	Rank-Order Function	[16, 27]
	Fractal Dimensions	[16, 17]
	Edge Detection	[14]
	Morphological Operations	[16, 17, 10]
	Eigen Filters	[16]
	Co-occurrence Matrix	[25, 16, 17, 28]
	Artificial Neural Network	[21, 29, 13, 30, 20, 35, 26, 31, 32, 33, 34, 36, 14, 37, 7]
	Decision Tree Logic using Artificial Neural Network	[35]
	Optimal Filter Design	[21]
Spectral Approaches	Fourier Transforms	[38]
	Gabor Filters	[10]
	Wavelet Transforms	[24, 39]
Model Based Approaches	Gauss MRF Model, Poisson's Model, Model Based Clustering	[24, 40]

III. RESEARCH APPROACH

Textile industry plays major role in everyone's life. Quality inspection is an important aspect of all industries while producing the quality product. There are lot of problems faced by the apparel industries such as:

- Inline inspection in the textile industry
- Fabric shrinkage
- Estimation of the porosity of textile fabrics.
- Measuring of thread densities
- Unavailability of quality monitoring tools.
- Inferior quality of raw materials
- Operations in a critical zone of the garment.
- Sewing on bias cut.
- **Defect detection and classification.**

Among all defect detection and classification plays a very significant role in producing quality product at spinning, weaving and processing side. There are lot of methods and algorithms has been developed on this but still research is going on for better method.

Machine vision automated inspection system for textile defects has been in the research industry for long-time [30]. Numerous techniques have been developed to detect fabric defects. The characterization of real fabric surfaces using their structure and primitive set has not yet been successful. Therefore on the basis of nature of features from the fabric surfaces, the proposed approaches have been characterized into three categories; statistical, spectral and model-based. However, random textured images cannot be described in terms of primitives and displacement rules as the distribution of gray levels in such images is rather stochastic. Therefore, spectral approaches are not suitable for the detection of defects in random texture materials [17]. The Fourier Transform (FT) has the desirable properties of noise immunity and enhancement of periodic features. Chan and Pang [40] used the Fourier analysis for fabric defect detection. The Fourier Transform is an analysis of the global frequency content in the signal, it is not able to localise the defective regions in the spatial dependency[17] in the spatial dependency into Fourier analysis through the windowed Fourier transform becomes the well-known Gabor transform, which can be achieving optimal localisation in the spatial and frequency domain[10]. In [41], Gabor filters are designed on the basis of the texture features extracted optimally from a non-defective fabric image by using a Gabor wavelet network (GWN). A major difficulty of this method is how to determine the number of Gabor channels at the same radial frequency and the size of the Gabor filter window in the application [17, 34, and 38]. Model - based texture analysis methods are based on the construction of an image model that can be used not only to describe texture, but also to synthesize it. Model-based approaches are particularly suitable for fabric images with stochastic surface variations [17, 34]. An important assumption in statistical approach is that the statistics of defects free regions are stationary, and these regions extend over a significant portion of inspection ages [17, 34].Co-occurrence matrix technique is based on different grey level configurations in a texture fabrics [16, 24] This co-occurrence technique can be computationally expensive for the demands of a real-time defect inspection system. The number of gray levels is usually reduced in order to keep the size of the co-occurrence matrix manageable [34]. Texture properties can be extracted by using several bi-dimensional transform such as Discrete Cosine Transform (DCT), Discrete Sine Transform (DST), Discrete Hadamard Transform (DHT), Karhunen–Loeve Transform (KLT) and Eigen filtering [17,34 30, 38]. This back propagation based neural network coupled with the DCT technique can lead to outstanding results for classification of various fabric defects[11].The defect detection approaches [14, 40] using edge detection are suitable for plain weave fabrics imaged at low- resolution. The difficulty in isolating fabric defects with the noise generated from the fabric structure results in high false alarm

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

rate and therefore makes them less attractive for textile inspection [34]. The Cross Correlation is used for locating features in one image that appear in another and therefore provides a direct and accurate measure of similarity between two images [16, 17 and 38]. Randomly textured backgrounds do not correlate well and demonstrate a limitation of this approach [17, 36]. Use of gray level thresholding enables to detect high contrast defects. The defect detection can be effective even when web is covered by a fine and complex pattern [16, 17, 22 and 24. The fabric inspection system that uses thresholding, proposed by Stojanovic et al. [42], gives high detection rate with good localization accuracy and low rate of false alarm. Detecting defects morphologically on spatially filtered images of fabrics produces better results [16, 17, 10 and 38], particularly when the fabric is fine and contains defect of small size. Thus the morphological operations are only performed on aperiodic images defects, unlike the case in [43] where the entire structure of thresholder fabric image was utilized. Neural networks are one of the fastest most flexible classifier used for fault detection due to their non-parametric nature and ability to describe complex decision regions [17, 34, 38 and 45]. A new approach for the segmentation of local textile defects using feed-forward neural network (FFN) and also a new low-cost solution for the task web inspection using linear neural network is presented in [35]. Loganathan and Girija [46] have used back propagation neural network, with fuzzy logic, to achieve the classification of eight different kinds of fabric defects along with defect-free fabric. The real-time implementation of defect segmentation scheme using FFN is computationally costly. Although the real time computational complexity of SVM is also similar, but do not suffer from the problem of local minimum and is computationally simple to train [17 45]. Habib [20], presented a novel hybrid model through integration of genetic algorithm (GA) and neural network to classify the type of garment defects. Experimental results for real fabric defect detection, shows the usefulness of the three intelligent techniques and they further stated that NN has a faster performance. Online implementation of the algorithms showed they can be easily implemented and may be adapted to industrial applications without great efforts. Their experimental result shows that this method can effectively detect defects and classify the types of defects with high recognition correct rate [17, 34, 38 and 45].

The over-all objective of this research it to develop an automatic defect detection and classification tool based on computer vision and artificial neural network decision tree classifier using resilient backpropagation algorithm. This application is robust alternative to traditional human visual systems. The achievement of this goal means that the following other sub-objectives have been consequently achieved:

1. Available literature is reviewed that addresses the fabric defect categorization and classification and also various technologies available for the defect detection and classification.
2. In this research we will deal mainly with two challenges i.e. defect detection and defect classification process.
3. Development of a fabric defect map will be generated with the help of industry experts to determine the major/minor/critical defects in knitted and woven fabrics which should be considered during the pre-processing step.
4. Acquisition or capturing of a sufficiently large fabric database of images in TIF/JPEG format with and without defects at different resolution levels for detection and classification process will be stored.
5. Suitable procedure using a software package (SCILAB or MATLAB) to implement the proposed technique (Computer Vision with Artificial Neural Network) will be developed.
6. Developing a methodology to extract defect features from various fabrics using various image processing techniques.
7. Identifying and optimizing the main parameters from the defective images which affects the defect detection process
8. Firstly the training of the images are done on the simulated fabric containing the chosen major defects, to understand the behaviour of the frequency spectrum, determine and optimize the most important detection parameters.
9. Designing of a computerized software to identify and classify the various defects using neural network decision tree classifier algorithm.
10. A computer demonstration of a sequence of steps from the pre-processing through the final detection
11. Test and verify the success of the technique using real plain fabric samples containing the same simulated defects.
12. Design and development of a prototype to examine the technique in real-time (during the production of the fabric on the weaving machine) that is the main object of this system.
13. Knowledge base for the expert system to provide online adaptive capabilities is efficiently generated.

IV. PROPOSED FRAMEWORK

This proposed methodology mainly focuses on the combination of image processing and artificial neural networks in textile industries research arena.

The main motive of this proposed method is to develop an economical automated system for fabric defect detection and classification for texture defects in textile industry. The purpose of the developed system is to reduce the labour cost, time, increase the productivity of the products and as well as increase accuracy in the inspection process and propose a better method. The location, size and image of the defect are recorded in the system. After the inspection process, the product will be graded in terms of severity and the detailed report will be generated. In addition to the use of standard image processing functions for enhancing and modifying the digital image, the paper will describe the techniques from artificial neural network for classifying the knitted and weaving defects.

A. Proposed Model:

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

In this work, a novel approach for defect detection and classification based on neural network decision tree classifier is presented. Based on the research, the proposed system design is divided into five parts. The first part of the defect recognizer focuses on the acquisition of image; second part involves the processing and normalization of the images to detect the faults; in the third part the image is filtered or noise is removed by using adaptive median filtering techniques into binary image by restoration and threshold techniques; during fourth part features are extracted from the pre-processed image; and in the fifth part the extracted features are given an input to the neural network decision tree classifier for further defect detection and classification. This whole software component is implemented using MATLAB. The block diagram for the proposed framework is given in Figure 1.

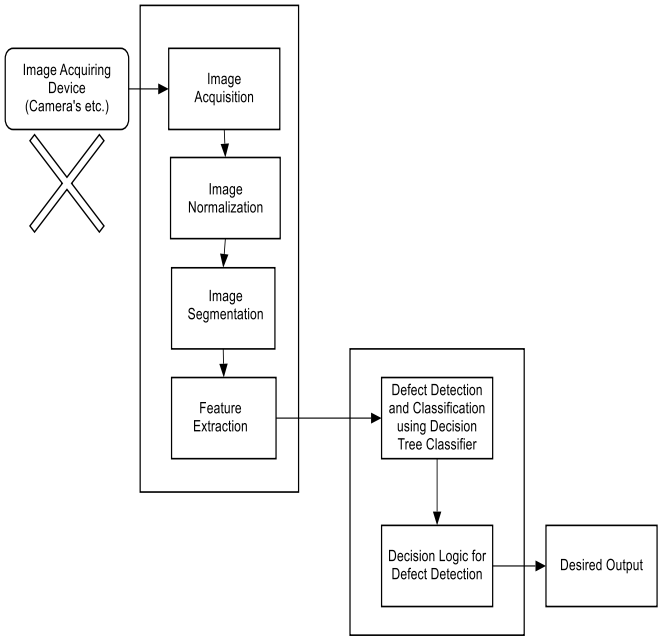


Figure 1. Computer Vision and Artificial Neural Network Framework for Defect Detection and Classification for texture Images

MATLAB is a high-level language and creates user interactive environment for numerical computation, visualization, and programming. MATLAB analyse data, develop algorithms, and create models and application. HDF, JPEG, PCX, TIFF, BMP, XWB are the image formats that can be used in MATLAB for processing. A brief overview of the process of MATLAB simulation for the method proposed is shown in figure 2.

B. Methodology:

1. **Image Acquisition:** The digital analysis of two-dimensional images of fabric is based on processing the image acquirement, with the use of the high resolution camera and computer. The most important parameter used in the image acquisition is the resolution. Either the size of one pixel or the number of pixels per inch can refer the resolution of an image. The lower the image resolution, the less information is saved and higher resolution means more information is saved but larger memory size is required to store. The scanning of fabric images begins from 300 dpi resolution because human vision is approximately 300 dpi at maximum contrast. The scanned image is stored in 'tiff /jpeg' or in grayscale format.

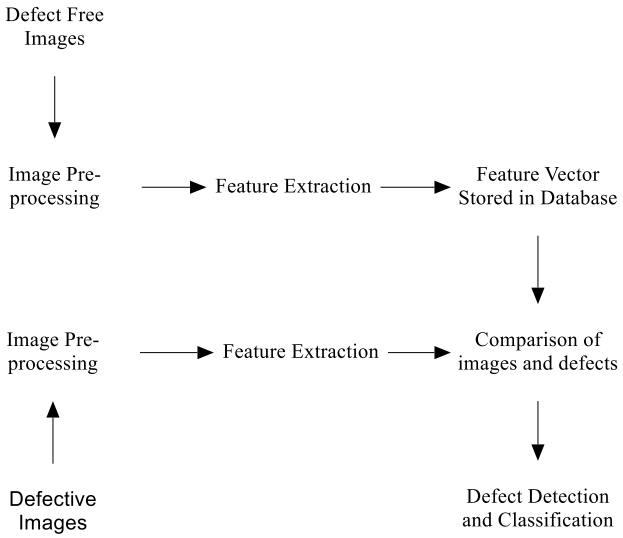


Figure 2. MATLAB Software Simulation Flow

2. **Image Normalization:** The acquired image may contain noise; noise is the result of errors in the image acquisition that result in pixel values. Noise reduction, filtering and thresholding is the process of removing noise from the image by using various image pre-processing techniques. In averaging filter, replace each pixel by the average of pixels in a square window surrounding this pixel. Larger window can remove noise more effectively but also blur the details/edges. The filtered image is converted into binary image, then the area of binary image is calculated

3. **Image Segmentation:** Image segmentation is typically used to locate an objects and boundaries (lines, curves, etc.) in images. Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Several methods [39] can be used to segment the defect, i.e., detect the defect on the image, ranging from simple segmentation methods (e.g., thresholding) to more advanced methods that combine background subtraction.

4. **Feature Extraction:** In this process the histogram of an image is drawn. The feature area is extracted from the binary image. This binary image is used to calculate the following attributes:
   - The area of the faulty portion: calculates the total defected area of an image.
   - Number of objects: uses image segmentation to calculate the number of labels in an image.
   - Shape factor: distinguishes a circular image form a noncircular image.

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

- Height of the overall defect windows
- Width of the overall defect window
- Ratio of total defect area to the overall window area.
- Number of defect occurs in the overall defects window.
- The ratio of the smallest defect area over the largest defect area.

These various attributes are used as input sets to adapt the neural net through training set in order to recognize and classify expected defects

5. **Defect Detection and Classification using ANN Decision Tree Classifier:** A neural network decision tree classifier is designed for classifying the type of defects. The various features describe in feature extraction technique is used as an input to the NN classifier. Initially a simple threshold value is calculated to identify the critical defects from the major and minor defects. A 3-layer Neural Network is designed for this purpose. During the training phase, the training data is fed into to the input layer. We are able to classify the defects with the attributes obtained from feature extraction technique. ANN contains four to six computing units in input layer, twelve to 25 computing units in hidden layer and six to 10 computing units in output layer .Each computing units in the output layer corresponds to each defect type. The data is propagated to the hidden layer and then to the output layer. Similarly each node in output layer gets input from all the nodes from hidden layer, which are multiplied with appropriate weights and then summed. The target output values are those that we attempt to teach our network. The error between actual output values and target output values is calculated and propagated back toward hidden layer. This is called the backward pass of the back propagation algorithm. The error is used to update the connection strengths between nodes, i.e. weight matrices between input-hidden layers and hidden output layers are updated. Figure 4 shows the block diagram of decision tree classifier and the way how it will identify the critical, major and minor defects of knitted and woven fabrics.

6. **Decision Logic for Defect Classification:** A decision tree is constructed based on the output extracted from the neural network classifier to detect and classify the various defects. Figure 3 shows the steps required in building of an automatic defect detection and classification system.

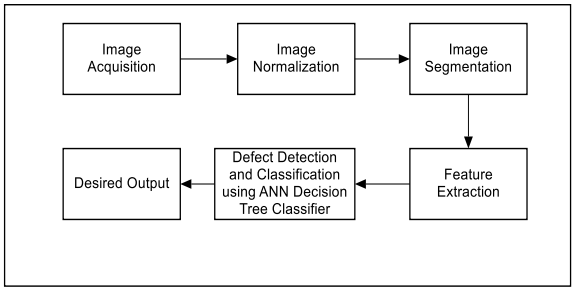


Figure 3. Steps required in development of Automatic Defect Detection and Classification of Texture Fabric Defects

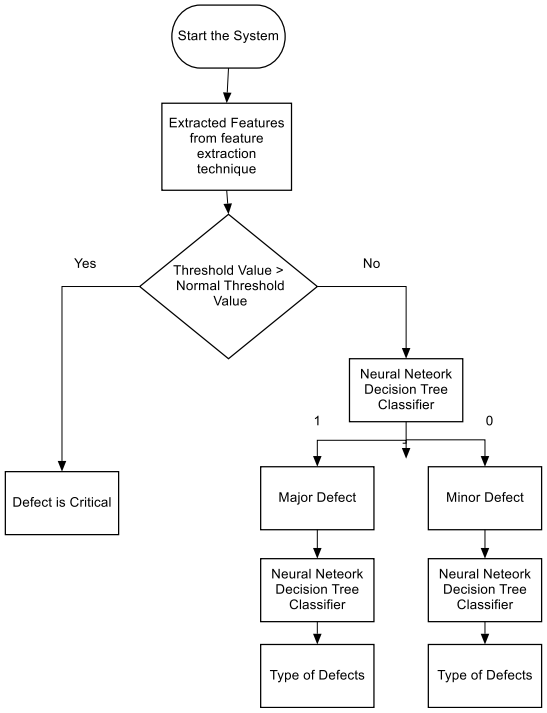


Figure 4. Block diagram of Decision Tree Classifier

The study result shows that the proposed method is feasible in textile production factories for defect detection and classification and can achieve success rate up to 90-95%. Figure 5 shows the detailed block diagram of defect detection and classification for texture defects in textile industry. Neutral network are one kind of the best classifiers for defect detection due to their non-parametric nature and ability to describe complicated decision region. In this proposed work neural network decision tree classifier is trained for classifying the defect types of textile fabrics.

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

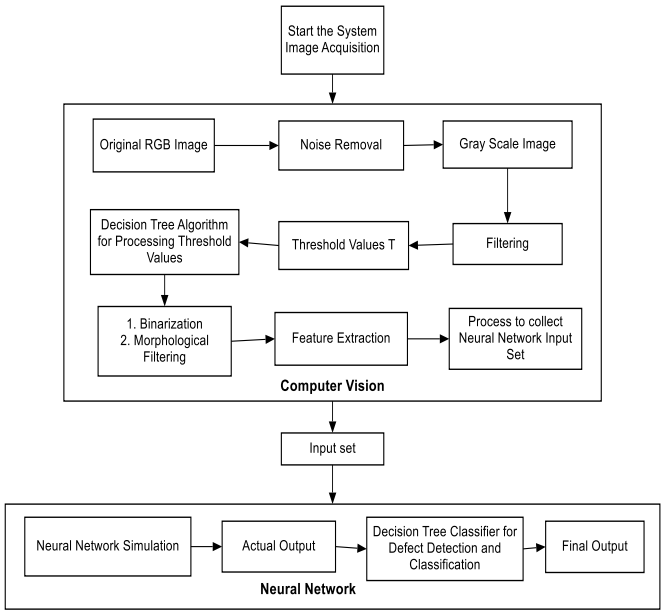


Figure 5. Block Diagram of Defect Detection and Classification for Texture Defects in Textile Industry.

V. DISCUSSION AND OBSERVATIONS

The automatic defect detector and classifier captures fabric images by acquisition device (digital camera) and passes the image to the computer. Initially by using image processing techniques defect detector normalizes the image and filtered it with adaptive median filtering. The number of connected components and their region property area with bounding box is calculated. Taking the value of area as threshold the image is converted into binary image. The texture features are obtained from the binary image which act an input to the neural network. The input layer consists of 4-6 neurons, hidden layer consists of around 25 neurons and output layer consists of 5-10 neurons. The number of neurons at output layer will represent the number of defects in the fabric. The neural network uses Log sigmoid algorithm as transfer function [34]. Mean of sum of squares of the network weights and biases is used for performance function mention in equation 1 and 2. The three layer neural network will be trained using the following function:

$$T = \{(X_i, D_i)^N_{\Sigma=1} \quad (1)$$

Where X_i = input vector of i^{th} example
Di = desired (target) response of i^{th} example
N = Training set size

Given the training sample T, the need is to compute the parameters of the neural network so that the actual output y_i of the neural network due to x_i is close enough to di for all i in a statistical sense. For example, we may use the Mean-Square Error (MSE) as the index of performance to be minimized.

$$E(n)=\frac{1}{N}\sum_{i=1}^{N}(D_i-y_i)^2 \quad (2)$$

The performance of the system can be evaluated by using different fabric images with defects and without defects.

After the integration of the defect detection algorithms and the artificial neural network classification algorithms to the final system. The proposed approach will be tested on the experimental platform. The results which can be generated from this approach are summarized in table (IV).

TABLE IV. RESULTS OF EXPERIMENTS

Parameter	Value
Number of Defects	10-12 Defects
Classification Accuracy with present of Noise	80-85%
Classification Accuracy without Noise	90-93%
Maximum Inspection Speed	120m/min
Width of Testing Fabric	1 m
Space Inspection Resolution	$1mm^2$
Overall Accuracy for Detection of Defect	90-95%

VI. CONCLUSION

From the above reviewed literature, it can be resolved that various defect detection and classification techniques have been deployed to find out the defects from various fabrics. Firstly different computer vision techniques were employed individually to find out the defects. This was not helpful as it lacks in defect detection accuracy, computational time increase and results were not accurate. Later on various artificial intelligence methods were combined with image processing techniques to overcome the above loop holes in research. Some of the techniques are KNN, Bayesian, SVM, Radial basis function, Gabor wavelet, Feed forward neural network algorithm, PNN etc. The core ides of these methodologies along with their drawbacks / critics were discussed. . In order to identify the formation and nature of the defects, it is important to accurately localise the defective regions rather than classifying the surface as a whole. This can help us to classify the defects and for further studies.

The proposed work can be used to detect and classify the defect for the low scale industries. In this research we have used image processing technique with neural network decision tree classifier to identify the several defects like hole, scratch, missing yarn, slub etc. We can achieve success up to 90%-94% to identify the defects. To obtain the better results we can combine both statistical and spectral approach together.

REFERENCES

1. J. Tou, R. Gonzalez, "Pattern recognition principles", Massachusetts, USA: Addison-Wesley Publishing Company, 1981.

2. YH Zhang., WK Wong, "An intelligent model for detecting and classifying color-textured fabric defects using genetic algorithms", The Elman Neural Network. Textile Research Journal, vol. 81, October 2011, ISSN no. 17 1772-1787, 2011.

3. Henry Y.T. Ngan, Grantham K.H. Pang, Nelson H.C. Yung, "Automated fabric defect detection – a review", vol. 29, issue 7, Science Direct June 2011, Page no. 442 – 458.

4. S. Priya T, Ashok Kumar, Dr. Varghese Paul, "A novel approach to fabric defect detection using digital image processing", Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011).

5. Ajay Kumar, "Neural network based detection of local textile defects", Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

6. G.Revathy, P.Vidhyalakshmi, "defect classification in fabric web material using labview". Volume 63– No.18, International Journal of Computer Applications (0975 – 8887), February 2013.

7. Md. Atiqul Islam, Shamim Akhter, Tamnun E. Mursalin, M. Ashraful Amin, "A suitable neural network to detect textile defects". ICONIP 2006, Part II, LNCS 4233, pp. 430 – 438, Springer-Verlag Berlin Heidelberg 2006.

8. Atiqul Islam, Shamim Akhter, Tumnun E. Mursalin, "Automated textile defect recognition system using computer vision and artificial neural networks". Proceedings of World Academy of Science, Engineering and Technology Volume 13 May 2006 ISSN-1307-6884.

9. Dr.R.S.Sabeenian, M.E.Paramasivam, P.M.Dinesh, "Detection and location of defects in handloom cottage silk fabrics using MRMRFM & MRCSF". Vol.2.No.2, International Journal of Technology And Engineering System, Jan –March 2011.

10. K.L. Mak, P. Peng, K.F.C. Yiu, "Fabric Defect Detection Using Morphological Filters", Image and Vision Computing 27 (2009) 1585–1592.

11. Aziz M.A, Haggag A.S, Sayed M.S, "Fabric defect detection algorithm using morphological processing and DCT", 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 2013, IEEE.

12. Md. Tarek Habib, M. Rokonuzzaman, "Distinguishing feature selection for fabric defect classification using neural network". Vol. 6, No. 5, Journal of Multimedia, October 2011.

13. Ali Javed, Mirza Ahsan Ullah, Aziz-ur-Rehman, "Comparative analysis of different fabric defects detection techniques", I .J. Image, Graphics and Signal Processing, 2013, page 40-45 Published Online January 2013 in MECS (http://www.mecs-press.org/).

14. Chi-ho Chan, Hugh Liu, Thomas Kwan, Grantham Pang. "Automation technology for fabric inspection system", Dept. of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong.

15. Dorrity J, Vachtsevanos G. and Jasper W, "Real-time fabric defect detection and control in weaving processes", National Textile Center, Annual Report, August, 1995, pp. 143-152.

16. Kumar A, "Computer vision-based fabric defect detection-A survey", Vol. 55, Issue 1,IEEE, Transactions on Industrial Electronics, , 2008, pp. 348-363.

17. Mahajan P.M, Kolhe S.R. and Pati P.M, "A review of automatic fabric defect detection techniques", Volume 1, Issue 2, Advances in Computational Research, ISSN: 0975–3273, 2009, pp.18-29.

18. J. Lewis Dorrity, George Vachtsevanos, Warren Jasper, "Real-time fabric defect detection and control in weaving processes", Project No. G94-2.

19. Chien H.-T, Sheen S.-H, Lawrence W. P, Razaziant K, and Raptis A. C. "On-Loom real-time noncontact detection of fabric defects by ultrasonic imaging", Review of Progress in Quantitative Non-destructive Evaluation (18B) , Plenum Press, New York, 1999, pp. 2217-2224.

20. Md. Tarek Habib, Rahat Hossain Faisal, M. Rokonuzzaman, "Feasibility of genetic algorithm for textile defect classification using neural network", Department of Computer Science and Engineering, Prime University, Dhaka, Bangladesh.

21. Kumar A. and Pang G, "Defect detection in textured materials using optimized filters", Vol. 32, No. 5, IEEE Transactions on Systems Man and Cybernetics, October 2002, pp. 553-570.

22. Dr. R.S. Sabeenian, M.E. Paramasivam. "Defect detection and identification in textile fabrics using multi resolution combined statistical and spatial frequency method". This work has been sponsored by the AICTE, Under Research Promotion Scheme from 2009-2012.

23. Yundong Li, Jingxuan Ai, Changqing Sun. "Online fabric defect inspection using smart visual sensors", www.mdpi.com/journal/sensors. Sensors 2013, 13, 4659-4673; ISSN 1424-8220.

24. Xie X, "A review of recent advances in surface defect detection using texture analysis techniques", Vol. 7, no. 3, Electronic Letters on Computer Vision and Image Analysis, 2008, pp. 1-22.

25. Sharda P. Jadhav, M.S.Biradar, "Fabric defect detection based on set of geometric features for neural network", Proceedings of 18th IRF International Conference, 11th January 2015, Pune, India, ISBN: 978-93-84209-82-7.

26. Thilepa R, "A paper on automatic fabrics fault processing using image processing technique in MATLAB, signal & image processing", Vol. 1, No. 2, An International Journal (SIPIJ), , December 2010, pp. 88 – 99.

27. Kar Seng Loke, "An approach to textile recognition", www.intechopen.com, Monash University Sunway Campus Malaysia.

28. Dr. G. M. Nasira, P.Banumathi, "Automatic defect detection algorithm for woven fabric using artificial neural network techniques", Vol. 2, Issue 1, International Journal of Innovative Research in Computer and Communication Engineering, January 2014.

29. Md. Tarek Habib, Rahat Hossain Faisal, M. Rokonuzzaman, Farruk Ahmed. "Automated fabric defect inspection: A survey of classifiers", Vol.4, No.1, International Journal in Foundations of Computer Science & Technology (IJFCST), January 2014.

30. Saduf, Mohd Arif Wani. "Comparative study of back propagation learning algorithms for neural networks", Volume 3, Issue 12, International Journal of Advanced Research in Computer Science and Software Engineering, , December 2013 ISSN-2277.

31. Dr.G..M. Nasira, P. Banumathi, "Plain woven fabric defect detection based on image processing and artificial neural networks", volume 6 No. 4, International Journal of Computer Trends and Technology, Dec 2013 ISSN: 2231-2803 Page 226.

32. Dr. G.M. Nasira, P. Banumathi, "Fabric defect identification system using statistical approach and artificial neural techniques", Research Journal of Applied Sciences (9):619-623, 2014, ISSN: 1815-932X Medwell Journals, 2014.

33. Tamnun E Mursalin, Fajrana Zebin Eishita, Ahmed Ridwanul Islam. "Fabric defect inspection system using neural network and microcontroller", Journal of Theoretical and Applied Information Technology © 2005 - 2008 JATIT.

34. Rashmi Mishra, Ms. Dolly Shukla, "A survey on various defect detection", Volume 10, Number 13 International Journal of Engineering Trends and Technology (IJETT) – Apr 2014.

35. Dr G.M.Nasira, Dr B.Nagarajan, Mrs P.Bhanumathi, "Artificial neural network based defect identifier", Volume: 4,| Issue : 12, Dec 2014, ISSN - 2249-555X.

36. Chi-Ho Chan, Pang, G.K.H, "Fabric defect detection by fourier analysis", Vol 36, Issue 5 Industry Applications, IEEE Transactions on, ISSN: 0093-9994.

37. Xuezhi Yang, Grantham Pang, Nelson Yung, "Discriminative training approaches to fabric defect classification based on wavelet transform", The University of Hong Kong.

38. Abdel Salam Malek, "Online fabric inspection by image processing technology". A thesis presented to the University of Haute Alsace in fulfilment of the thesis requirement for the degree of Doctor of Philosophy Year: 2012.

39. Jagrti Patel, Meghna Jain, Papiya Dutta, "Detection and location of defects fabrics using feature extraction technique", Vol. 5 Issue 3, International Journal of Emerging Trends in Engineering and Development, September 2013.

40. Chi-ho Chan, Grantham Pang, "Fabric defect detection by fourier analysis", Proceedings of the 1999 IEEE Industry Applications Society Conference Phoenix, Arizona, USA, October 3-7, 1999.

41. Mak K. L.; Peng P. "Robotics and Computer- Integrated Manufacturing", 24(3), 359-369, 2008.

42. Stojanovic R, Mitropulos P, Koulamas C, Karayiannis Y, Koubias S, and Papadopoulos G. "Defect detection and classification on web textile fabric using multiresolution decomposition and neural networks," Proc. IEEE Intl. Conf. Electronics Circuits and Systems, pp. 765-768, Sep.1999.

43. Yixiang Frank Zhang, Randall R. Bresee, "Fabric defect detection and classification using image analysis", vol. 65 no. 1 1-9, Textile Research Journal January 1995.

44. Dr. C.Loganathan, K.V.Girija, "Hybrid learning for adaptive neuro fuzzy inference system", Vol.2, Issue 11, Research Invents: International Journal of Engineering and Science,April 2013, Pp 06-13.

45. Ali Javed, Mirza Ahsan Ullah, Aziz-ur-Rehman, "Comparative analysis of different fabric defects detection techniques", An I.J. Image, Graphics and Signal Processing, 2013, 1, 40-45 Published Online January 2013 in MECS.

AUTHORS PROFILE

Shweta A Loonkar has received her B.E (Computer Science & Engineering) degree from University of Mumbai. She has completed her M. Tech (Computer Science) from Banasthali University, Rajasthan. She is Ph.D research scholar in Computer Engineering Department of MPSTME, NMIMS University. She is having around 10 years of experience in academics. She is working as a visiting professor at Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS. She has published the book System Programming and Compiler Construction for 3rd year students of computer science for Mumbai University. Her area of interests are Compiler Construction, Artificial Intelligence, Neural Networks, Fuzzy Logic and Image Processing.

Dr. Dhirendra S. Mishra has received his B.E (Computer Engineering) degree from University of Mumbai. He has completed his M.E. (Computer Engineering.) from Thadomal Shahani Engineering College, Mumbai, and University of Mumbai. He has completed his Ph.D. in 2012. He is working as an Associate Dean and Head of Computer Engineering Department of Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS University. www.nmims.edu/Engeneering , Mumbai, INDIA. He is life member of Indian Society for technical education (ISTE) www.isteonline.in & Member of International Association of Computer Science and Information Technology (IACSIT), Singapore, Membership Number. : 80337336, http://www.iacsit.org . He is Member of International Association of Engineers (IAENG), Membership Number: 106347, http://www.iaeng.org.He has 60 papers published in various National/International Conferences/Journals to his credit. His areas of interests are Image Processing- Image database, Image retrieval, Pattern recognition, Operating systems, Information Storage and Management.

Management System of Children with Autism

Elham Mohammed Thabit AbdAlameer
Karbala University / College of Science / Department of Computer Science
Iraq, Karbala

***Abstract*- Autism is one of the most prevalent chronic disease; it is a developmental disability that causes problems with communication and social interaction shared with repetitive demeanour. Some behaviors could occur before age three and appear through delays in diverse skills which develop from childhood to adulthood, even though behaviors may optimize over time. The disorder includes a large spectrum of symptoms, levels of impairment, and skills. It varies in severity from a handicap which somewhat limits an otherwise normal life to a devastating disability which may need institutional care. However, optimization of health care is probable to have a good effect on progress of disease, quality of life, and functional outcome. This paper will design and built a reliable system based Electronic healthcare technologies to care with autistic child. It is centralized database which contains the patients profiles; the system will enable the nurse to access the database and enter details of patients for subject them to a set of tests ; then these data will be held on a database and the system will identify the type of autism to determine the suitable treatment to each type. This proposed clinical system will serve the patient and health care providers by reducing healthcare costs and enable parents to help their autistic children by using the best way to treat them.**

Keywords- ASDs, Autistic children, Neurobehavioral, Clinical system, Ehealth care systems, Pediatricians.

1 INTRODUCTION

Autism is a spectrum of strongly related disorders with a common core of symptoms. It appears in early childhood, causing delays in several important areas of development like learning to speak, move, and interact with others. Most of autistic children have a learning disability, also known as mental retardation. There are many theories have been proven that the origins of autism may be caused by a genetic susceptibility to an environmental [1]. The children who have autism may look normal but their behavior could be downright difficult. The signs and symptoms of autism vary widely in individuals with autism , as do its effects. Some of children have only mild impairments, but others have more obstacles to overcome [2]. Generally, every child with autism shares problems with each other, at least in these three areas: Relating to others and the world around them, communicating verbally and non-verbally, and behaving and thinking flexibly[3].

There are different opinions among parents, physicians, and experts in Autism Spectrum about what causes this disease and how best to treat it, and they still don't know. Fortunately recent advances in technology have opened up the doors to enable scientists to tackle the issue. This paper will introduce the understanding of Autism Spectrum, its types, symptoms, causes and how best to treat it by using a clinical system which diagnosis and treats the children with ASDs. This system will manage this disease and help those children who need a careful care from their families or their carers

1.1 Autism Understanding

Autism is a neurodevelopmental disorder that typically involves delays and impairment in developmental language, behavior, and social skills. And it refers to people having dissimilar behaviors along a spectrum. It describes qualitative differences and weakness in reciprocal social interaction, combined with repetitive behaviours. According to the criteria defined in the international Statistical Classification of Diseases, Autism spectrum are diagnosed in children, young person and adults [4]. It is a lifelong disorder that has important impact on the child or young people and their parents or carers. Patients with Autism may have a deep sense of relief that others agree with their concerns and observations [5]. Diagnosis and the evaluation of needs could help an understanding the obstacles which face the patients and can open roads to support and services in social care and health services into voluntary organisations and make contact with other children and parents who have similar experiences; all of these can optimize the lives of the children and their carers. However, because of children special social problems, nowadays many of tools available to evaluate social skills, such as Ehealth Care Systems, and to ensure efficient system which serve the children with ASD, we need to determine what each individual needs

1.2 AUTISM TYPES AND SYMPTOMS

The Autism disorder ranges in severity from a handicap which impedes normal life to other disability may be devastating that may require rapid intervention and special care. Children with autism disorder have trouble communicating and they are very sensitive, they have difficult to understand the others think and feel . Also they may be affected quickly from influences that surround them like: touches, smells, sounds, and they sometimes find it difficult to express on the pain [6]. Basically, there are different types of ASDs have been determined by guidelines in the diagnostic manual (DSM - IV) of the American Psychiatric Association. And according to the CDC, the three major types of Autism Spectrum Disorders are: Asperger's syndrome, Pervasive developmental disorder, and Autistic disorder [7].

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

Children with autism have trouble picking up on subtle nonverbal cues and using body language; this makes it very hard for them to express themselves, facial expressions, and touch, and makes the reciprocal social interaction very difficult [8]. However, early signs of autism and symptoms in toddlers in autism like:

They don't make eye contact normally.
They don't respond to the sound of a familiar voice.
They don't use gestures to communicate.
They don't make noises to get your attention.
They don't track objects visually.
They don't respond to cuddling.
They don't imitate your facial expressions or movements.
They don't attract or play with other people.

1.3 Autism Causes

Autism is defined as a complex developmental disorder; experts believe that Autism appears in the first three years of a child's life as a result of a neurological disorder which has an effect on normal brain function, and affecting social interaction skills [9]. Genomic research discovered the that the children with ASDs a probably share genetic traits with individuals with attention- disorder, schizophrenia, or clinical depression [10]. However ASDs has no single known cause, and there are probably many causes may play a role like both genetics problems, as a result of a malfunction in the brain ; and environment factors, such as problems during pregnancy , viral infections, and other [11].

1.4 Autism Treatment

The main goals of treatment are to enhance the quality of child's life by minimizing the core ASDs features, promoting socialization, reducing maladaptive behaviors, and guiding and supporting parents So it is essential to find helpful services, education, and treatments for autistic children. Physicians have an key role in early recognition, evaluation of autism disorders, and also in chronic management of these disorders.

Treatment and care should take into account the needs of children, and their families and carers who care for them [12]. They should participate with health care professionals for the success the treatment process of their children. However, there are many treatments that can assist autistic children to learn new skills and overcome their disability. According to the National Institutes of Health treatments for autism disorders can involve the following: Behavioral management therapy, Educational therapies, Cognitive behavior therapy, Medication treatment. In addition other treatments and therapies that have been used for autistic children, such as: Speech and language therapy, The Picture Exchange Communication System, Music therapy, Vitamins and mineral supplements, and Massage therapy[13].

In the proposed system, the treatments for autistic children are include the chemo therapy and physical therapy as shown in system interfaces section.

2 The Proposed System Aims

The aim of the proposed system is to serve health care providers by procedure a set of tests to identify the type of autism, and make a suitable decision based on the collected data to identify the required treatment for each autism type. This system is used to manage and monitor autism disorders in autistic children and it is efficient due to reduce the costs of health and enhance the communication between a patient and providers of a health care.

2.1 Brief System Description

The scope of this work is to build a reliable system based Electronic healthcare technologies; the clinical database of management system of children with autism stores details of the users related to this system . It has four users which were identified as: doctor, nurse, admin , and technician; each one has collection of functions in side system. The nurse has the ability to register, login in, login out, add – edit patient account, research patient profile , and test patient state. While the doctor has ability to register, login in, login out, search patient profile, view the result of test and the treatment, print the medical prescription of treatment, and follow up the patient case during duration of treatment. Whilst the admin has ability to register, login in, login out, add, edit user account, and manage the system. Lastly the technician is able to register, login in, login out, and check system hardware and buck up the system.

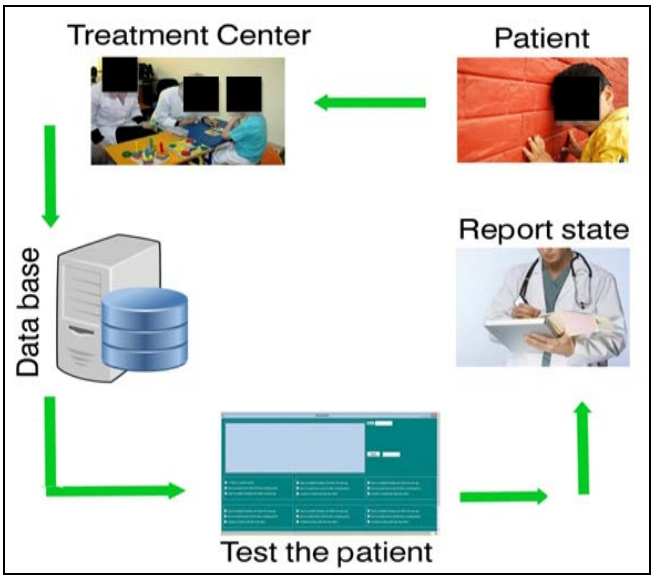


Figure 1: System Architecture for Management System of Children with Autism

2.2 MoSCoW Prioritisation

The technique of MoSCoW Prioritisation helps to produce a strong system [14]. It will apply in the Management System of Children with Autism. First of all, we will do several interviews with health care providers like doctor, nurse and families of patients to identify the functional requirements of

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

the clinical system. These requirements will identify according to the priorities as the following:

- Input staff information (Should).
- Add new Patient account (Must).
- Tests patient state (Must).
- Determine the type of disease (Must).
- Determine the treatment for each type of disease (Must).
- Print the medical perception of treatment (Should).
- List all appointments for the coming week (Should).
- List all staff (Should).
- Report on historical appointments (Could).

2.3 Data Flow Diagram

The data flow diagram (DFD) is a tool that shows the data flow through a system and the work and processing performed by that system. It is used to help understand the existing system and to represent the required system. The diagram as shown in figure 2 represents the external bodies sending and receiving information[15], [16].

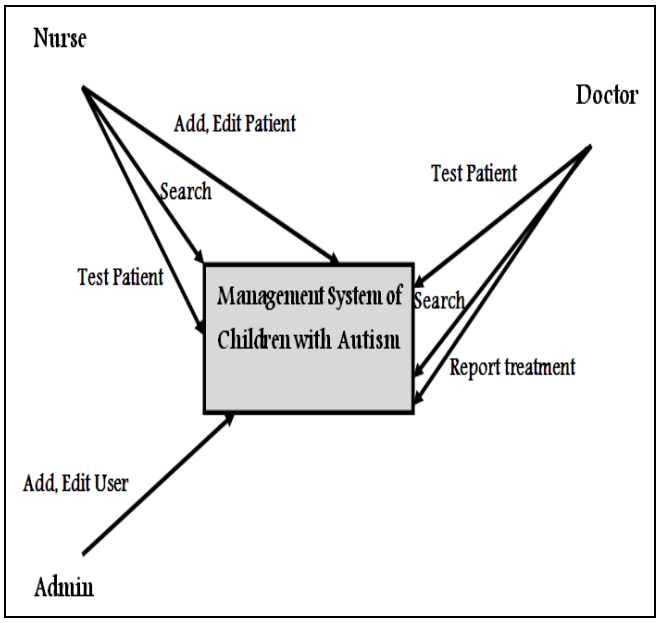


Figure 2: Data flow diagram of the system

2.4 User Involvement

During the development this system, it is essential to gather information about users who related to the system [17]. In Management System of Children with Autism four users were identified like: doctor, nurse, admin, and technician and each of them has several responsibilities and roles as shown in table 1.

Table 1: Roles of users

Actors	Roles
Nurse	- Register - Log in and log out - Add Patient account - Edit Patient account - Search patient account - Test patient state - Determine the type of disease
Doctor	- Register - Log in and log out - Search patient account - View the result of test and the treatment - Determine the treatment for each type of disease - Print the medical prescription of treatment - Follow up the patient case during duration of treatment
Administrator	- Register - Log in and log out - Add, Edit user account - Manage the system
Technician	- Buck up the system

3 Proposed System Design

The purpose of design phase is to investigate what the software will look like, both graphically and functionally. Its purpose is to create a technical solution which satisfies the system functional requirements [18]. In this phase, we will design the clinical database of the management system of children with autism. It considers very important phase, a well-designed database provides the correct information for the medical decision-making process to succeed in an efficient way. However, database design can be used to describe many different parts of the design of an overall database system.

3.1 Use Cases Models

Use Case model illustrates a set of possible scenarios related to a particular goal. Each use case represents an action the system is required to allow an associated user to achieve. It allows the definition of the system's boundary, and the relationship between the system and outside of the system [19]. In Management System of Children with Autism, there are four users in the use case diagram. Each user have many functions in the system as shown in figure 3.

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

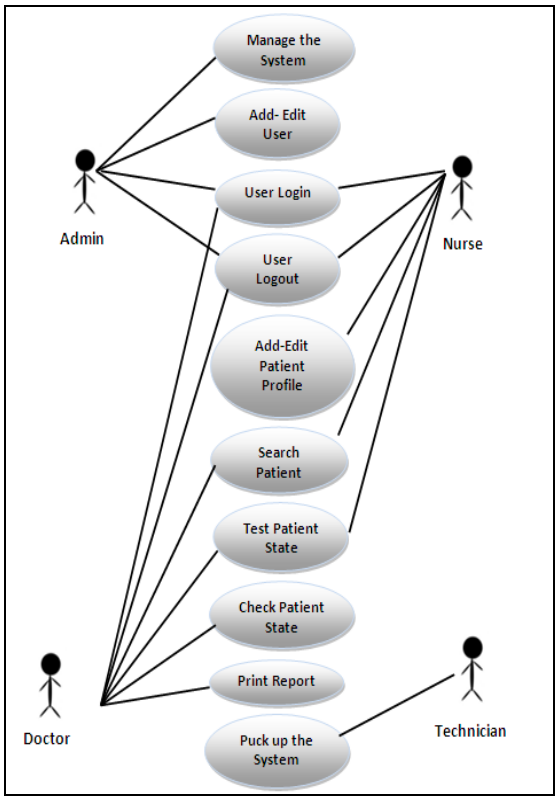


Figure 3: Use Cases Model

3.2 Implementation Phases

In the clinical database design of the Management System of Children with Autism, three main phases will be create: conceptual, logical and physical design.

The first stage describes the relation and the connectivity between all components of the system. While the logical data model consists of specified classes that will become tables which includes six tables as shown in table 2; then the attributes of tables will become fields, and the associations become relationships. Lastly, the final phase will translate the logical database into a physical database [20].

Table 2: Contents of Management System of Children with Autism

No.	Table Name	Description
1.	Patient	It contains information related to the patient
2.	Disease	It contains information about the disease and it's types
3.	Symptoms	It contains information about the symptoms of the disease
4.	Treatment	It contains information about the treatment of the disease
5.	physical therapy	It contains information about physical therapy of the disease
6.	Chemo Therapy	It contains information about chemo therapy of the disease

3.3 Used Tools

In Management System of Children with Autism, Microsoft visual studio 2010 and Microsoft sql server 2008 were used to be an easy yet powerful way to create management of system that actually interact with the users. Microsoft visual studio 2010 was used to design the interfaces of the system and implement the Programming code. While Microsoft sql server 2008 was used to create the database, its tables , and data and connect it to the system.

4 System Interfaces and Results

The interfaces and results that have been obtained through the implementation of Management System of Children with Autism will be showed in this section:

1. Login Interface: The first step is to login in to the system by admin and nurse by enter their own username and password as shown in figure 4; then the page as shown in figure 5 will appear which contains the database of the patient which includes two options: add new patient, and edit patient data.

Figure 4: Login interface

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

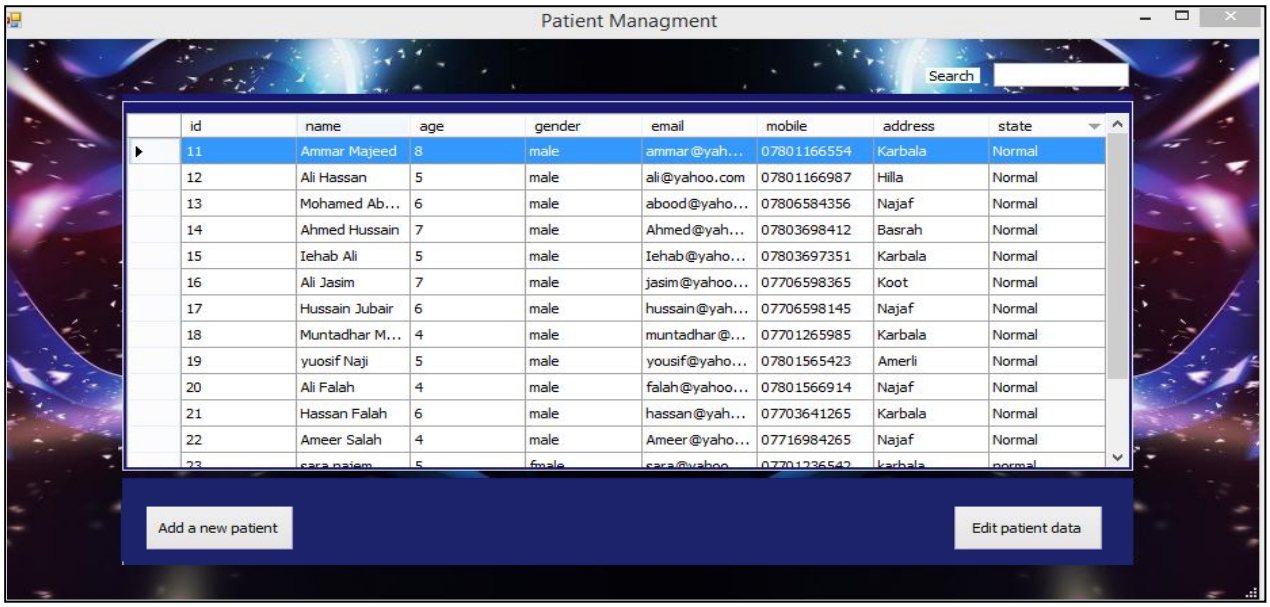


Figure 5: Database of the patient

2. Figure 6 shows the process of adding new patient account completed successfully. While figure 7 shows the process of updating also completed successfully.

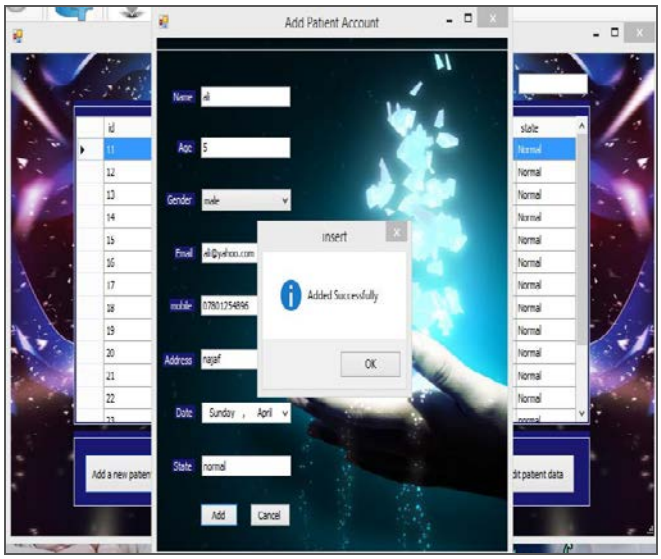


Figure 6: Added successfully

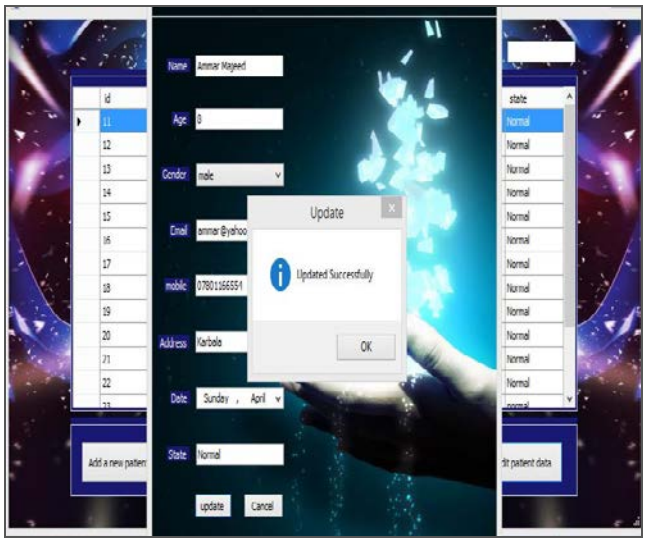


Figure 7: Updated successfully

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

3. Tests interface: when the nurse select option tests in the main page, the page of tests will appear as shown in figure 8 to identify the type of autism. While figure 9 shows the interface of the tests results and identify the type of disease. And Figure 10 shows the details of each type of autism disease.

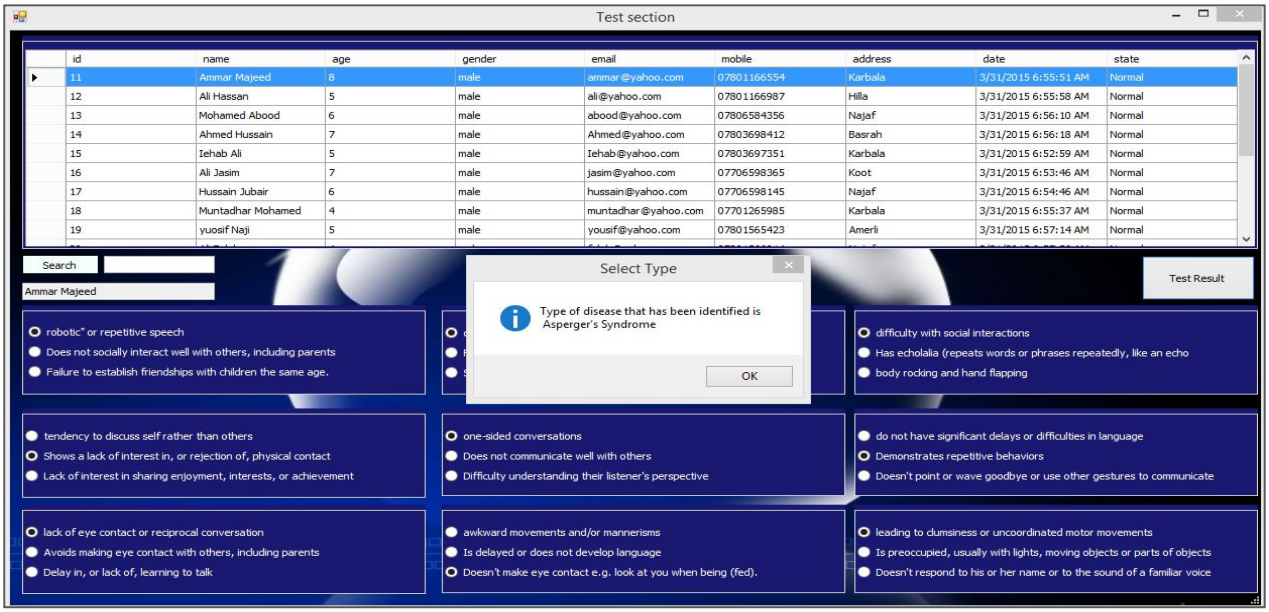


Figure 8: Tests interface

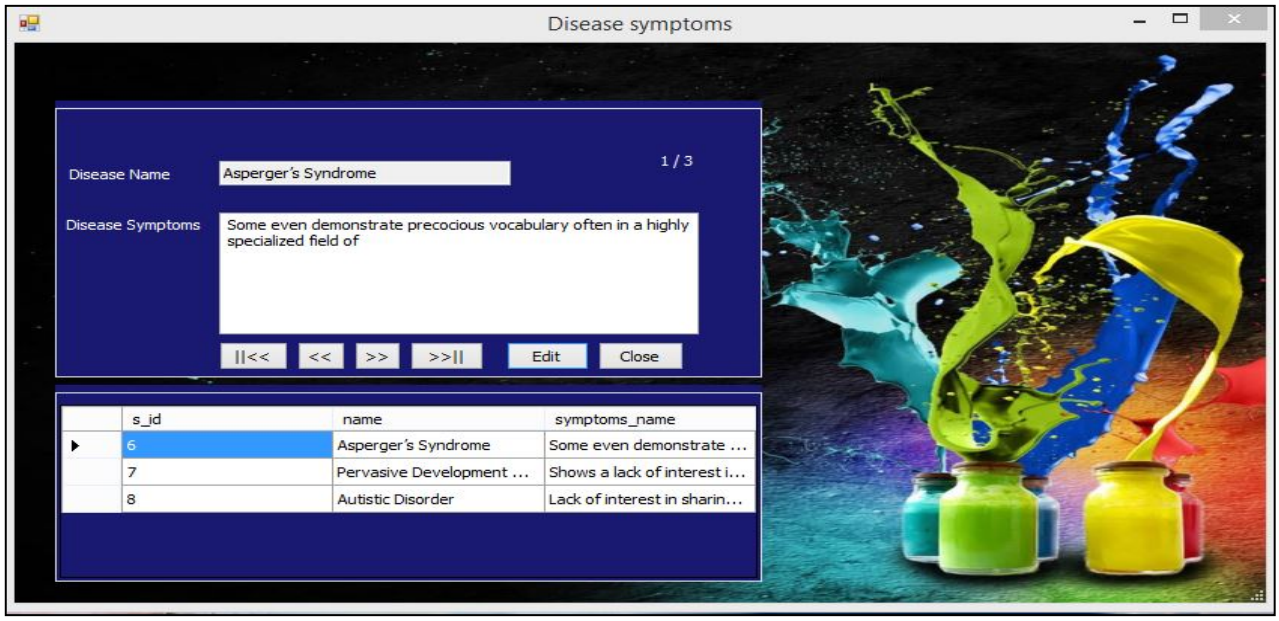


Figure 9: One of Autism types

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

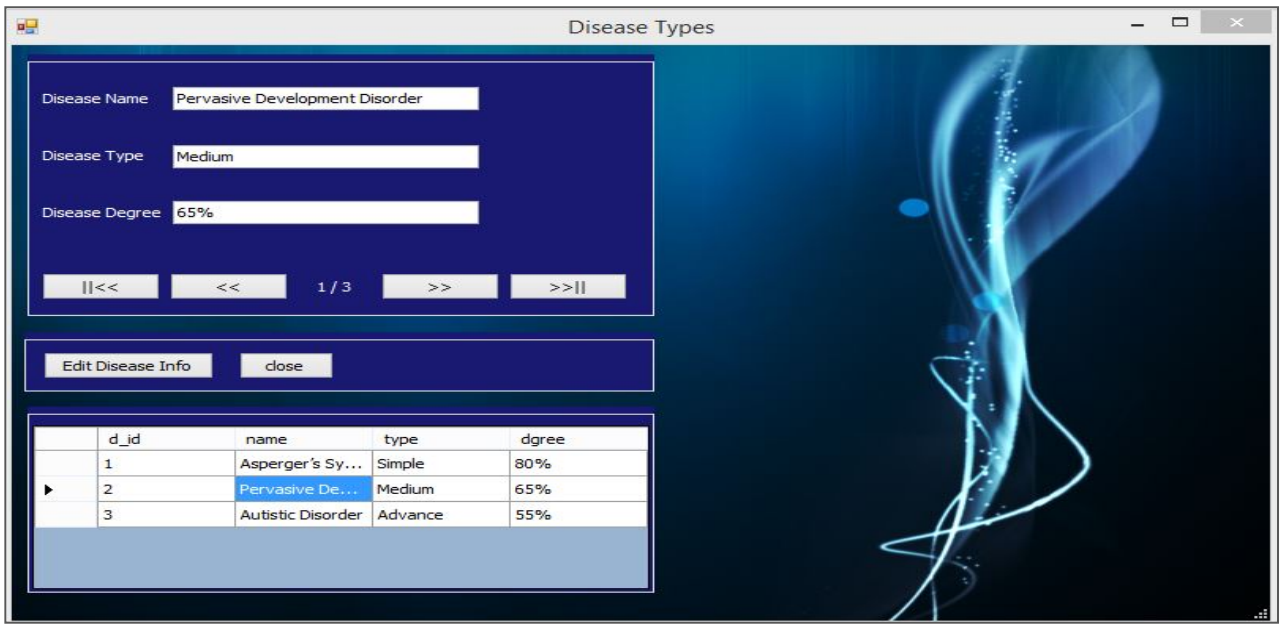


Figure 10: Details of each type of disease interface

4. Figure 11 shows the interface of treatment for each type of disease. While figure 12 illustrates treatment of Chemo Therapy. Lastly figure 13 illustrates the medical prescription of treatment which will print to the families patients.

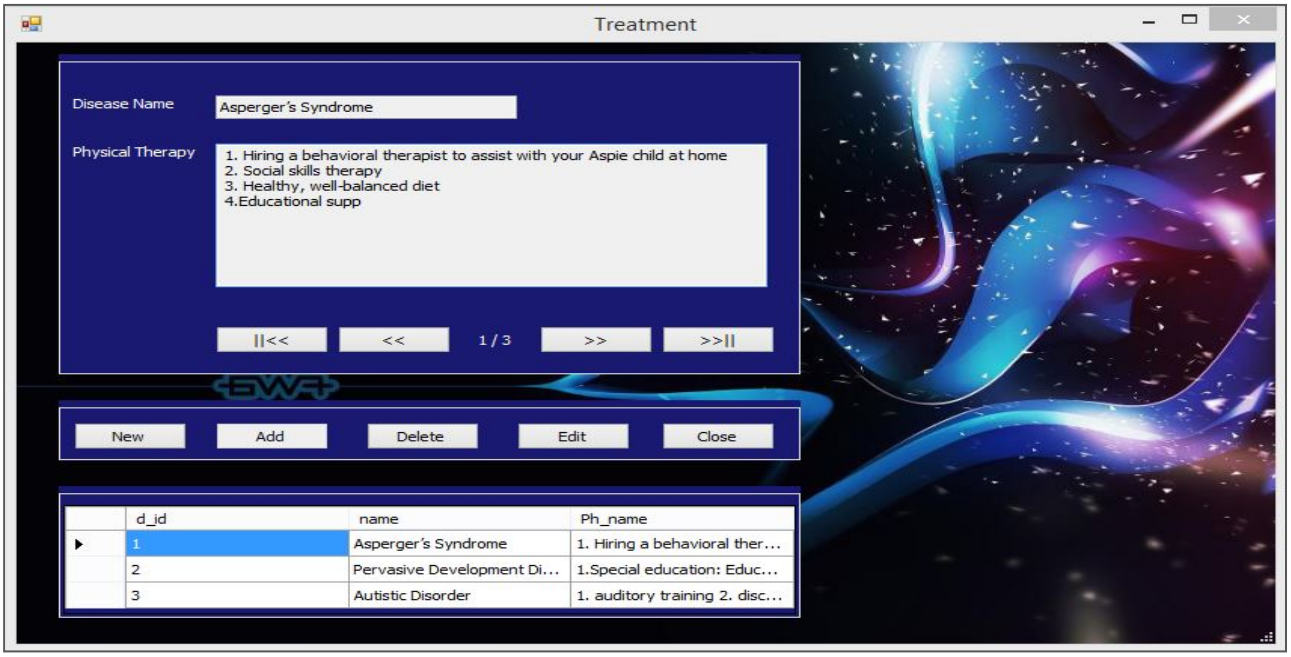


Figure 11: Treatment for each type of disease

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

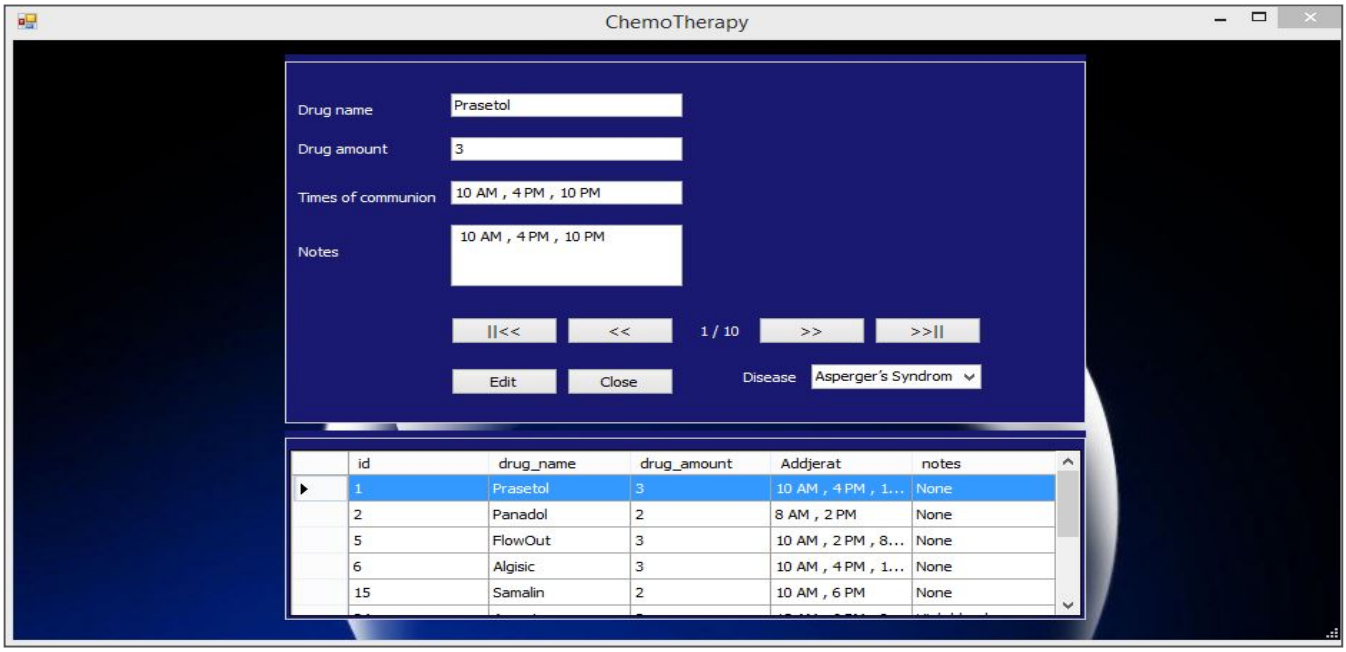


Figure 12: Treatment of Chemo Therapy

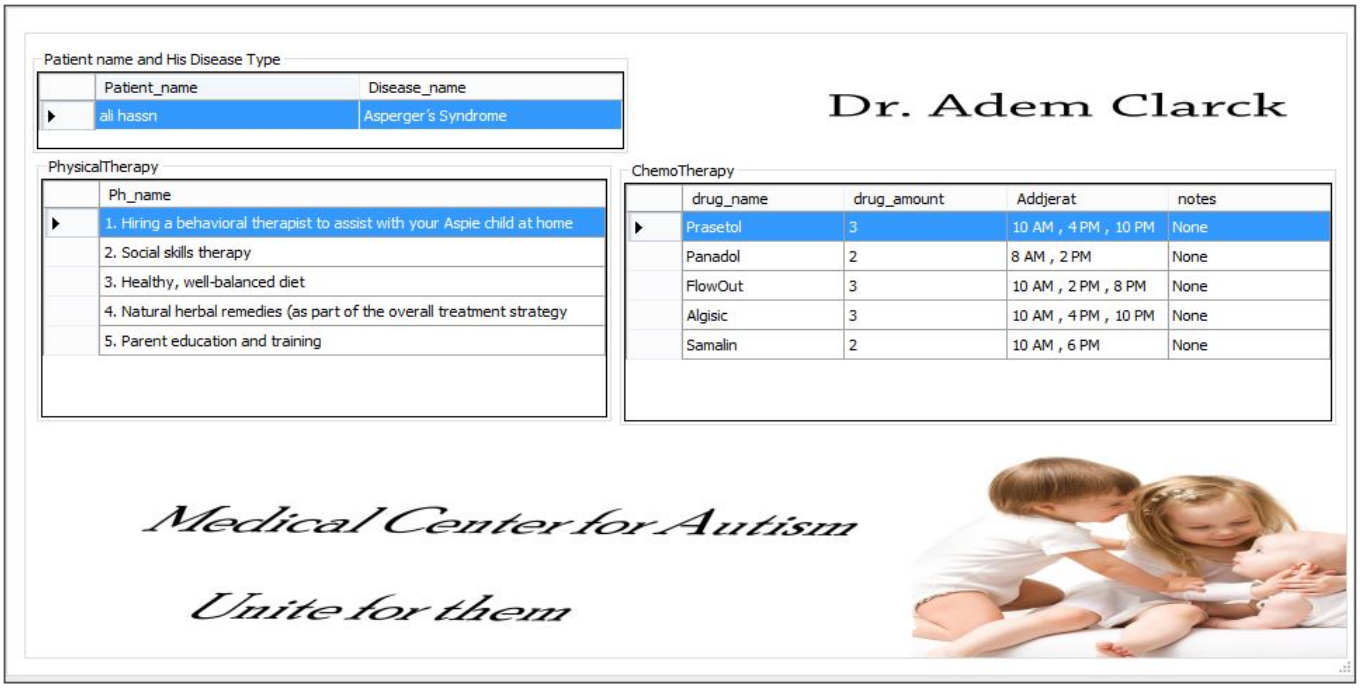


Figure 13 : Medical prescription

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

5 Conclusion and Future work

This section will focus on the conclusion and future work for Management System of Children with Autism.

5.1 Conclusion

Good communication between healthcare providers and autistic children is essential; It should be supported by evidence-based written information customized to the needs of the child and their parents or carers. This paper has been presented the design and implementation of a reliable system based ehealth care technologies for autism spectrum management care. The aim of this work was to provide a description to the basic building blocks involved in the Autism management system. It has been provided the autistic children with reliable ways to manage their disease even outside clinic doctor. The system produced a set of tests to identify the type of autism. And then it will allow the health care provider to make an informed healthcare decision based on the collected data to determine the required treatment for each type. Lastly, the recommendations of treatments are given to the parents of child by printed medical prescription.

However, after implementing this system; it will be capable of achieving the following:

- It is to enhance the provided services to patients by making their records available for doctor to follow up the case easily with less effort;
- Autism Spectrum management has dependability and consistency execute the required functions of software;
- It provides the best control of the patients status based on their test results.

Lastly, we can say when the autistic children find the assistance to meet their special needs, with availability the correct treatment plan, and a lot support they will educate and improve.

5.2 Future work

This section will discuss future improvements to Management System of Children with Autism by adding the additional functionalities to this system :

- Apply this system on the web;
- Enabling the patient to take appointment with the doctor before go to the health centre;
- Allow the families and carers who care for autistic children to access into the database to enter families patients requires and enable them to receive the latest developments and recommendations to follow up the medical advices by practitioners.

References

[1] James B.,Stephen M., and etal., "Advice for Parents of Young Autistic Children", (2004): Working Paper. Available at: http://www.autismtoday.com/AdviceForParents.pdf

[2] Akhila, T., Sharmin, D, "Dental Management and Behavioural Modifications of Children with Autism Spectrum Disorders (ASD)", EC Dental Science 1.1 (2015): 33-55. Available at: https://www.ecronicon.com/ecde/pdf/ECDE-01-00007.pdf

[3] Notbohm, E, " How well the family functions as a whole is just as important as how well the child with special needs is doing, Ten Things Every Child with Autism Wishes You knew", 2010. Available at: https://www.autismspeaks.org/sites/default/files/images/10_things.pdf

[4] NICE National Institute for Health and Care Excellence, "Autism diagnosis in children and young people, Recognition, referral and diagnosis of children and young people on the autism spectrum", 2011, http://www.nice.org.uk/guidance/cg128/resources/guidance-autism-diagnosis-in-children-and-young-people-pdf

[5] Dautenhahn, K., "Design Issues on Interactive Environments for Children with Autism", 2013, College Lane, Hatfield, Herts AL10 9AB, UNITED KINGDOM. Available at: https://www.autismspeaks.org/sites/default/files/images/10_things.pdf

[6] Online document, healthisall.in ,Understanding autism and helping those with special needs. Available at:
http://beta.healthisall.in/index.php?route=information/healthplusview&blog_id=66

[7] RESEARCH AUTISM, IMPROVING THE QUALITY OF LIFE, Types of Autism. Available at: http://www.researchautism.net/autism/types-of-autism [Last access January 1, 2015]

[8] Lindgren,S., Doobay, A., May 2011, "Evidence-Based Interventions for Autism Spectrum Disorders" http://www.uihealthcare.org/uploadedFiles/UIHealthcare/Content/Services/Center_for_Disabilities_and_Development/UCEDD/DHS%20Autism%20Interventions%206-10-11.pdf

[9] The Arc, Autism Spectrum Disorder (ASDs), 2012, Available at:
http://www.thearc.org/document.doc?id=3668

[10] Karen L., 2006, The Official Autism 101 E-Book: Autism Today, Available at: http://autismtoday.com/autism-101-ebook.pdf

[11] Scott M. Myers, and etal .,(2007), "Management of Children With Autism Spectrum Disorders", American Academy of Pediatrics, the Council on Children With Disabilities, http://www2.aap.org/pressroom/issuekitfiles/managementofchildrenwithasd.pdf

[12] TRIAD Social Skills Assessment, Assessing Children with Autism Spectrum Disorders, 1998, 2010 Vanderbilt TRIAD, Vanderbilt Kennedy Center. Available at: http://vkc.mc.vanderbilt.edu/assets/files/resources/tssamanual.pdf

[13] AMERICAN MUSIC THERAPY ASSOCIATION, Music Therapy As a Treatment Modality for Autism Spectrum Disorders, 2012. Available at:
http://www.musictherapy.org/assets/1/7/MT_Autism_2012.pdf

[14] UKESSAYS, DSDM Atern Approach And Agile Development Information Technology Essay. Available at:
http://www.ukessays.com/essays/information-technology [[Accessed 25 January 2015]

[15] Satzinger, J., Jackson, R., and etal., (2007), YSTEMS ANALYSIS AND DEDIGN IN A CHANGING WORLD, 4th ed. Available at:
http://faculty.kutztown.edu/tan/csc354/Datafiles/A.pdf

[16] Online document, Creating Date Flow Diagrams in Yourdon Notation, Available at:
http://www.philblock.info/hitkb/c/creating_data_flow_diagrams_yourdon.html

[17] Online document ,"User and Task Analysis for Interface Design", available at:

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

http://redish.net/books/item/43-user-and-task-analysis-for-interface-design

[18] Jarke M., Lenzerini M., etal., (2003), "Fundamentals of Data Warehouses". Springer-Verlag.

[19] Rosenberg D. and Stephens M., (2007), "Use Case Driven Object Modeling with UML: Theory and Practice", Apress.

[20] Connolly T. and Begg C., (2005), "Database Systems Apractical Approch to Design, Implementation, and Management", 4th Edition, addition-Wesley.

AUTHORS PROFILE

Elham Mohammed Thabit AbdAlameer

Awarded her B.Sc at Al-Mustansiriya University in Iraq, College of Science, Department of Computer Science in 1994, and M.Sc, at Kingston University London in UK, Faculty of Computing, Information Systems and Mathematics in 2010 respectively.

She is a lecturer at Department of Computer Science, College of Science, Karbala University, Karbala, Iraq.

Here research interests include: Information Systems, System Analysis , Database, and E- business.

Requirement Tracing using Term Extraction

Dr. Najla Al-Saati
Software Engineering Dept
College of Computer Sciences & Mathematics
Mosul, Iraq

Raghda Abdul-Jaleel
Software Engineering Dept
College of Computer Sciences & Mathematics
Mosul, Iraq

***Abstract*- Requirements traceability is an essential step in ensuring the quality of software during the early stages of its development life cycle. Requirements tracing usually consists of document parsing, candidate link generation and evaluation and traceability analysis. This paper demonstrates the applicability of Statistical Term Extraction metrics to generate candidate links. It is applied and validated using two datasets and four types of filters two for each dataset, 0.2 and 0.25 for MODIS, 0 and 0.05 for CM1. This method generates requirements traceability matrices between textual requirements artifacts (such as high-level requirements traced to low-level requirements). The proposed method includes ten word frequency metrics divided into three main groups for calculating the frequency of terms. The results show that the proposed method gives better result when compared with the traditional TF-IDF method.**

Keywords- Requirements Traceability; Traceability Analysis; Candidate Link Generation; Parsing; Term Extraction; Word Frequency Metrics.

I. INTRODUCTION

The traceability of requirements was introduces mainly to manage and document the life of requirements. Its major objective is to maintain the activities of critical software development, for instance, the assessment of whether a software system has satisfied its definite set of requirements, the verification that all requirements have been employ by the end of the lifecycle, and the analysis of the impact imposed by the proposed changes on the system [1].

It is usually essential to follow the changes of requirements all the way through the lifecycle of software. All requirements should be validated in and at the end of each phase of the lifecycle. Traceability matrices are usually constructed to show the satisfaction of requirements by the design [2].

Generating traceability links (or traceability matrices) is fundamental to many software engineering activities [3]. But it is a time consuming, error prone, and mundane process. Most frequently, traceability matrices are built manually. When an analyst tries to trace a high level requirement document to a lower level requirement specification, he may have to look through M x N elements, where M and N are the number of high and low level requirements, respectively. Keeping in mind that there are very few tools available to assist the analysts in tracing unstructured textual artifacts, and those require enormous pre-processing [2].

Verification and Validation (V&V) and Independent Verification and Validation (IV&V) are used to ensure that the right processes have been used to build the right system. That is why it must be verified that the agreed processes and artifacts are directing the development in each phase of the life-cycle, in addition to ensuring that all requirements have been implemented at the end of the lifecycle. A requirements traceability matrix (RTM) is necessary for both of these [4][5].

The automatic generation of traceability links requires Information Retrieval (IR) techniques to reduce the time needed to generate the traceability mapping [3].

Requirements tracing usually enclose: document parsing, candidate link generation, candidate link evaluation, and traceability analysis. There are two commonly used measures for evaluating candidate link lists: *recall* and *precision*. In candidate link evaluation, the analyst investigates the candidate links and determines those that are actual (true links), and those that are not (false-positives, bad links). To achieve this, the analyst visually inspects the text of the requirements to find out the meanings of the requirements, compare them, and decide based on his believes which meanings are adequately close. This decision is based on human judgment and tolerates all the advantages and disadvantages that are related to it [4][5].

When tracing is finished, reports are generated by the analyst stating the high level requirements that do not have children and the low level elements that do not have parents (traceability analysis) [4][5].

II. RELATED WORK

Many researchers have presented their work in requirement tracing during the last few years, such as:

In 2004 Hayes, et al. [5] designed RETRO to support the IV&V analyst in requirements tracing to find and evaluate candidate links.

Also in 2004, Sundaram, Hayes, and Dekhtyar [6] studied a mixture of IR methods used to solve the requirement traceability problem. They found that existing IR methods can be used in automating the generation of candidate links with minimal modification. And that the analyst's feedback information can considerably improve requirements tracing.

By 2006 Hayes, Dekhtyar, and Sundaram [4] inspected the efficiency of information retrieval methods in automating the tracing of textual requirements. They found that feedback from analyst improves final results via objective measures.

In 2007, Sundaram [2] assisted analysts in the traceability links generation process with information retrieval techniques for improving the quality of the generated links in addition to time saving.

Finally in 2010, Sundaram, et al. [3] stated that Information Retrieval techniques have been shown to aid in the automated generation of links through reduction of the time used in generating the mapping of traceability.

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

Researchers have successfully used techniques such as Latent Semantic Indexing (LSI), Vector Space Retrieval, and Probabilistic IR.

III. REQUIREMENT TRACING

Requirements tracing is defined as the ability to describe and follow the life of a requirement, in both a forward and a backward direction, through the whole systems life cycle [2].

During the process of requirement gathering, the analyst has to clarify customer needs, conduct feasibility studies, specify a solution, and cross validates the specifications [7].

In large-scale projects, it is quite possible to miss or misinterpret some of the recognized requirements. More than 80% of the failures in large-scale mission-critical projects are caused by undetected problems in the early phases of the software development lifecycle [8]. An unobserved problem at the start of the project can continue all the way through to the deployed product; and becoming a latent defect or latent error [7].

Two sets of documents are typically created in the early phases of any software project:

- **Software Requirements Specification SRS**
  It is defined as "documentation of the essential requirements (i.e., functions, performance, design constraints, and attributes) of the software and its external interfaces. The software requirements are derived from the system specification [7]. SRS is a "binding contract among designers, programmers, customers, and testers," it includes different design views or paradigms for system design [9].
- **Software Design Description SDD**
  The design activity is used to identify the components of the software design and their interfaces from the Software Requirements Specification. The principal artifact of this activity is the Software Design Description (SDD) [9]. It is a "representation of software created to facilitate analysis, planning, implementation, and decision making". It is used as a medium for communicating software design information, and may be viewed as a blueprint or model of the system [7].

At the end of a requirements tracing process, a requirements traceability matrix (RTM) is generated [2]. RTM acts as a tool for indicating the way that the design and implementation elements deal with requirements throughout the whole software development lifecycle [7].

IV. INFORMATION RETRIEVAL (IR) FOR REQUIREMENTS TRACING

Information retrieval (IR) is the process of discovery documents relevant to an information request in a collection of documents, usually a search query [7].

The main issue in IR is the determination of relevant documents in document collections given user-specified information needs. Most IR methods work by converting each document in the collection into a mathematical representation to capture the information content of the document, after that a comparison is conducted with similar representations of user information needs (queries). Nearly all IR methods are keyword-based: the document and query representations contain information regarding the importance of particular keywords found in the document [10]. There is a broad array of keyword-based retrieval models meant for document collections. The Boolean model is the simplest: a representation of a document is a Boolean vector identifying the keywords found in the document. A *Vector model* broadens the Boolean model by correlating each term in the document representation with a *weight* that signifies its understood importance to the document collection [11].

Documents and queries are represented as a vector of keyword weights. Formally, let $\boldsymbol{V} = \{\boldsymbol{k_b}..., \boldsymbol{k_N}\}$ be the vocabulary of a given document collection. Then, a vector model of a document d is a vector $(\boldsymbol{w_b}, ..., \boldsymbol{w_N})$ of keyword weights, where $\boldsymbol{w_i}$ is computed as in Eq. (1) [10] [11].

$$w_i = tf_i(d) \cdot idf_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where
$tf_i(d)$ is the term frequency of the ith keyword in document d ,
idf_i is the *inverse document frequency* of the ith term in the document collection.

Term frequency is the number of term occurrences in the document and is usually normalized. The Inverse document frequency is computed using Eq. (2) [10][11].

$$idf_i = log_2\left(\frac{n}{df_i}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Where
df_i is the total number of documents containing the ith term in the document collection, and
n is the size of the document collection.

The term significance is judged by how often this term is located in the document and by how discriminating the term is. That is, less frequent terms have more important presence for the document. A user query is also converted into a similar vector $\boldsymbol{q}=(\boldsymbol{q_1},\ldots,\boldsymbol{q_N})$ of term weights. In this model, given a document vector $\boldsymbol{d}$ and a query vector $\boldsymbol{q}$, the similarity between them is computed as the cosine of the angle between vectors $\boldsymbol{d}$ and $\boldsymbol{q}$ in the N-dimensional space as in Eq. (3) [10][11].

$$sim(d,q) = \cos(d,q) = \frac{\sum_{i=1}^{N} w_i \cdot q_i}{\sqrt{\sum_{i=1}^{N} w_i^2 \cdot \sum_{i=1}^{N} q_i^2}} \quad \ldots\ldots\ldots\ldots(3)$$

V. EMPLOYD FILTERS

In this work, four filters are introduced to generate candidate link lists with relevance higher than one of the predefined levels: 0, 0.05, 0.2, and 0.25. This filtering acts as an assessment of the quality for the candidate link list. Having two candidate link list, say list X and list Y, with the same recall and precision, in that case if the true links show up at the top of list X compared with list Y, then obviously list X have preference to list Y from the analyst standpoint [2].

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

VI. MEASURING THE EFFICIENCY

To evaluate the efficiency of IR techniques, recall (*R*) and precision (*P*) are used as the primary measures. recall measures if a method succeeded in finding all the high-low level requirement pairs that trace to each other, while recall indicates the number of additional pairs found by the method that do not trace to each other[6].

The computation of recall is done by dividing the total number of relevant retrieved documents by the total number of relevant documents in the complete collection, as in Eq.(6) [12].

$$R = \frac{\#\text{of\_relevant\_retrieved}}{\#\text{\_relevant\_in\_collection}} \quad \text{.....(6)}$$

The precision is calculated as the total number of relevant retrieved documents divided by the total number of retrieved documents, as shown by Eq.(7) [12].

$$p = \frac{\#\text{of\_relevant\_retrieved}}{\#\text{\_retrieved}} \quad \text{.....(7)}$$

VII. TERM EXTRACTION

Term extraction forms an important issue in natural language processing; its goal is to extract sets of words with precise meaning in a collection of text. More than a few linguists considered these terms to be the base semantic unit of language. Automating term extraction comprises machine translation, automatic indexing, building lexical knowledge bases, and information retrieval [13].

Both supervised and unsupervised techniques have been used in earlier investigations to extract and distinguish terms. Nearly all researches aimed at locating the most significant set of terms from a domain corpus, to be precise, the set of superficial representations of domain concepts that better symbolize the domain for a human expert [14].

Term frequency in a corpus is a basic statistical property. This may then be compared to the frequency of the term in other corpora, such as balanced corpora or corpora from other domains. Basic frequency counts are integrated to compute co-occurrence measures for words. Co-occurrence measures are employed to estimate the propensity for words to appear together as multi-word units in documents, and to estimate the likelihood that units on either side of a bilingual corpus correspond under translation [15].

Term extraction can be used in this work to solve two issues:

- Finding high and low level requirements to create a common vocabulary. This is carried out using Statistical approaches, where all the terms are placed in a common vocabulary without any repetition.
- Using Statistical Term Extraction Metrics to calculate term weighting instead of TF-IDF in information retrieval.

VIII. STATISTICAL TERM METRICS

In this work, ten standards metrics are proposed each as a measure instead of that used in the TF-IDF method, which was mentioned in Eq.(1). These metrics are divided into three main groups as explained in the following subsections [16].

Through the next subsections the following notations are used to symbolize equations: tf_{ij} is the frequency of term *i* in each document *j*, *N* is the size of corpus. w_i is the weight of term *i*.

A. Term Frequency Based

The majority of term extraction algorithms base their results on some computation concerning term frequency.

1) Corpus Term Frequency

This metric is a solely term frequency metric, calculated over the entire corpus. It focuses on words that appear more often, except that it consequently favors large documents. Eq.(8) shows this calculation [16].

$$w_i = \sum_{j=1}^{N} tf_{ij} \quad \text{..... (8)}$$

2) Logged Term Frequency

Logarithms are considered as powerful modifiers of data, as they can reduce the range of values in a set. Logarithms are used to reduce the range of terms in any given document. This dampens the data, decreasing the distribution of frequencies as in Eq.(9) [16].

$$w_i = \sum_{j=1}^{N} \ln(tf_{ij} + 1) \quad \text{..... (9)}$$

3) Document Term Frequency

The maximum term frequency in a document is a unique metric, where the words that appeared most within their respective document are selected instead of summing together all the term frequencies. This is normalized, so as not to penalize words in short documents. This may provide new terms to the vocabulary by finding terms that appear often in one document, but not in any of the others. It favors unevenly distributed word frequencies, the calculation is done according to Eq.(10) [16].

$$w_i = \max_{1 \le j \le N} tf_{ij} \quad \text{.....(10)}$$

B. Normalization Based

Term normalization forms a standard metric for information retrieval; it is carried out by dividing the frequency of a term by the total number of terms in a document. When each document is normalized, the effect of size is removed, and each term frequency will form a percentage of another characteristic of the document, such as the document's term count [16].

1) Document Terms Counts

The widespread normalization of a document is carried out by dividing a term's frequency by the number of terms in a document [16]. Assuming Tj to be the total term count in document *j*, w_i is calculated as in Eq.(11).

$$w_i = \sum_{j=1}^{N} tf_{ij}/T_j \quad \text{..... (11)}$$

2) Document Maximum Frequency

In this metric, the term frequency is divided by the most frequent term in a document, and the results are then summed up. The most frequent word gets a score of one

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

for the document for which it is the most frequent term, in addition to any score it obtains by occurring in other documents. This has a similar effect to normalization because the score given to a term from any single document will not be greater than one, but the scores resulting from each document will be different than the scores after standard normalization. The weight *w* of term contributions is a ratio of the term frequency to the most common term P_j, rather than the frequency to the document size. Eq.(12) depict this [16].

$$w_i = \sum_{j=1}^{N} tf_{ij}/P_j \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

a) Document Maximum Frequency & Term Average Frequency

This metric also employs normalization according to the most frequent word in the document P_j, but here the average frequency that term *i* appears across as documents in the corpus is subtracted from Eq.(12). This is calculated as in Eq.(13) [16].

$$w_i = \left(\sum_{j=1}^{N} tf_{ij}/P_j\right) - \frac{\sum_{i=1}^{N} tf_{ij}}{N} \;\ldots\ldots\ldots\ldots\ldots (13)$$

3) Corpus Maximum Frequency

The previous maximum frequency normalization technique can be further explored by using the most frequent term in the corpus. Being fixed, the corpus's most common term is a constant P_c. Results should be similar to the results of term frequency, if not exactly the same [16]. This metric is sometimes called corpus relativized. Eq.(14) shows this calculation.

$$w_i = \sum_{j=1}^{N} tf_{ij}/P_c \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (14)$$

a) Corpus Maximum Frequency & Term Average Frequency

Referring back to the previous metric, the normalization was based on the most frequent term in the corpus, this metric is corpus relativized minus the average of TF as in Eq.(15) [16].

$$w_i = \left(\sum_{j=1}^{N} \frac{tf_{ij}}{P_c}\right) - \frac{\sum_{i=1}^{N} tf_{ij}}{N} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (15)$$

C. Inverse Document Frequency

The inverse document frequency measures desire words appearing in very few documents. It is used employed frequently in indexing; this is due to the fact that indexed documents in the corpus are in general varied, so a term that appears in few documents is a good identifier for those documents. Inverse Document Frequency together with term frequency assists in selecting words that occur repeatedly, but only in few documents [16].

1) The TF-IDF

Here, the term frequency is multiplied by the number of documents in the corpus, which is divided by the number of documents (n_i) that contain the term [16]. Eq. (16) shows the weight calculation.

$$w_i = \sum_{j=1}^{N} tf_{ij} * N/n_i \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (16)$$

2) Logged IDF

This is similar to the TD-IDF measure, but here the term frequency is weighted more highly. The logarithm decreases the range of IDF values as in Eq.(17) [16].

$$w_i = \sum_{j=1}^{N} (tf_{ij}) * \ln\left(\frac{N}{n_i}\right) \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (17)$$

IX. DATASETS

This work is validated using two NASA open source datasets. Both MODIS and CM-1 datasets are used here to assess the utilized techniques of IR. The MODIS dataset consists of 19 high level and 49 low-level requirements, where the CM-1 dataset contains 235 high-level requirements and 220 design elements. A manual tracing was done for both datasets for verification; these are referred to as “answer sets” or “theoretical true traces”. There were 41 and 361 true links found for the MODIS and CM-1 datasets, respectively [6].

X. EXPERIMENTAL RESULTS

Term Extraction is presented in this paper, along with a discussion of the Preprocessing techniques that are commonly used. First, the documents are parsed using the Statistical approach, stop words (words such as ‘the’ and ‘of’) are removed, and each remaining term is stemmed using Porter’s algorithm [17], the term frequency is computed using ten word frequency metric rather than TF-IDF. In this paper the vector space model is used for Information Retrieval.

The four filters were used together with the metrics described previously using MODIS and CM1 datasets. The results are compared with those found by Sundaram et al. [6].

A. First Dataset (MODIS) with Filters (0.2 and 0.25)

In this section experiments are done using the MODIS Dataset and filters (0.2 and 0.25). Table (I) and (II) show the results of running the ten metrics for each filter. It was found that:

- **Filter 0.2**, recall value for all metrics improved, the best value was (68.2) achieved by the Document Term count metric and is labeled with (*) in Table (I). Best Precision is (23.7) in Term Frequency – Inverse Document Frequency metric.
- **Filter 0.25**, Recall improved for nearly all metrics except for Document Term Frequency, best value was (68.2) achieved by the Document Term count metric and is labeled with (*) in Table (II). Best Precision is (21.6) in Term Frequency – Inverse Document Frequency metric.

TABLE 1.RESULT OF METRICS IN MODIS DATASET WITH FILTER (0.2)

Format	Term Weighting	*Recall*	*Precision*
XML[6]	**TF_IDF**	19.5	21.6
Statistical	**Corpus Term Frequency**	**65.8**	13.5
Statistical	**Logged Term Frequency**	**65.8**	14.2
Statistical	**Document Term Frequency**	**24.3**	7.6
Statistical	**Document Terms Counts**	**68.2***	17.1
Statistical	**Document Maximum Frequency**	**63.4**	16.5
Statistical	**Document Maximum Frequency and Term Average Frequency**	**65.8**	17.0
Statistical	**Corpus Maximum Frequency**	**65.8**	13.7
Statistical	**Corpus Maximum Frequency and Term Average Frequency**	**65.8**	13.4
Statistical	**Term Frequency – Inverse Document Frequency**	**34.1**	**23.7***
Statistical	**Logged Inverse Document Frequency**	**65.8**	14.0

** best value*

TABLE II.RESULT OF METRICS IN MODIS DATASET WITH FILTER (0.25)

Format	Term Weighting	*Recall*	*Precision*
XML[6]	**TF_IDF**	19.5	32.0
Statistical	**Corpus Term Frequency**	**65.8**	16.0
Statistical	**Logged Term Frequency**	**63.4**	16.4
Statistical	**Document Term Frequency**	17.0	7.6
Statistical	**Document Terms Counts**	**68.2***	19.3
Statistical	**Document Maximum Frequency**	**63.4**	19.5
Statistical	**Document Maximum Frequency and Term Average Frequency**	**63.4**	19.6
Statistical	**Corpus Maximum Frequency**	**65.8**	16.1
Statistical	**Corpus Maximum Frequency and Term Average Frequency**	**65.8**	16.0
Statistical	**Term Frequency – Inverse Document Frequency**	**19.5**	21.6*
Statistical	**Logged Inverse Document Frequency**	**65.8**	18.7

** best value*

B. Second Dataset (CM1) with Filters (0 and 0.05)

Here, experiments are done using the CM1 Dataset and filters (0 and 0.05). Table (III) and (IV) show the results of running the ten metrics for each filter. It was found that:

- **Filter 0**, best Recall is (98.6) in Document Term Frequency and Term Frequency – Inverse Document Frequency metrics. Best Precision is (1.0) for all metrics as in Table (III).

- **Filter 0.05**, best Recall is (95.2) in Term Frequency – Inverse Document Frequency metric, Best Precision is (1.1) in Document Term Frequency, Term Frequency – Inverse Document Frequency and Logged Inverse Document Frequency metrics as in Table (IV).

TABLE III.RESULT OF METRICS IN CM1 DATASET WITH FILTER (0)

Format	Term Weighting	*Recall*	*Precision*
XML[6]	**TF_IDF**	97.8	1.5
Statistical	**Corpus Term Frequency**	97.7	1.0
Statistical	**Logged Term Frequency**	**98.0**	1.0
Statistical	**Document Term Frequency**	**98.6***	1.0
Statistical	**Document Terms Counts**	97.5	1.0
Statistical	**Document Maximum Frequency**	97.5	1.0
Statistical	**Document Maximum Frequency and Term Average Frequency**	97.5	1.0
Statistical	**Corpus Maximum Frequency**	97.7	1.0
Statistical	**Corpus Maximum Frequency and Term Average Frequency**	97.5	1.0
Statistical	**Term Frequency – Inverse Document Frequency**	**98.6***	1.0
Statistical	**Logged Inverse Document Frequency**	**98.3**	1.0

** best value*

TABLE IV.RESULT OF METRICS IN CM1 DATASET WITH FILTER (0.05)

Format	Term Weighting	*Recall*	*Precision*
XML[6]	**TF_IDF**	92.2	4.3
Statistical	**Corpus Term Frequency**	86.9	1.0
Statistical	**Logged Term Frequency**	87.5	1.0
Statistical	**Document Term Frequency**	**93.0**	1.1
Statistical	**Document Terms Counts**	86.1	1.0
Statistical	**Document Maximum Frequency**	86.9	1.0
Statistical	**Document Maximum Frequency and Term Average Frequency**	86.9	1.0
Statistical	**Corpus Maximum Frequency**	86.9	1.0
Statistical	**Corpus Maximum Frequency and Term Average Frequency**	86.1	1.0
Statistical	**Term Frequency – Inverse Document Frequency**	**95.2***	1.1
Statistical	**Logged Inverse Document Frequency**	**92.7**	1.1

** best value*

In MODIS dataset, the **Recall** measure for both filters (0.2 and 0.25) showed better result for all metrics when compared to [6] except for Document Term Frequency in filter 0.25. Using the **Precision** measure, only Term Frequency – Inverse Document Frequency showed better results in filter(0.2), in filter 0.25 all of metrics showed less result than [6].

In CM1 dataset best value obtained in **Recall** measure was by using filter 0 and metrics (Logged Term Frequency, Document Term Frequency, Term Frequency – Inverse Document Frequency and Logged Inverse Document Frequency), which showed better results than [6], in filter 0.05 the Document Term Frequency, Term Frequency – Inverse Document Frequency, Logged Inverse Document

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

Frequency were better than [6].In **Precision** all of metrics showed less result than [6].

In this work, focus was on improving recall at the cost of precision because high-recall, low-precision lists of links appear to be more preferable than high-precision, low recall links[4][5]. That is due to the fact that humans may be better at deciding if a specific pair of links in the list is a match than at finding new pairs of links in the document [5].

XI. CONCLUSIONS AND FUTURE WORK

In this paper, the effectiveness of information retrieval methods in automating the tracing of textual requirements was examined. Ten metrics were evaluated and it was found that better recall can be achieved when compared to TF-IDF.

In this work, the vector space model was adapted for each of the metrics, in addition to the Statistical format. Porter Stemming Algorithm was applied using two open source datasets (MODIS and CM1).

Future work can carry on in several directions, such as the use of another technique in Information Retrieval (IR), as well as the vector space model. More methods can be sought to be employed other than term extraction to enhance results. Other datasets can also be used in this area.

REFERENCES

[1] X. Zou "Improving Automated Requirements Trace Retrieval Through Term-Based Enhancement Strategies", Dissertation for the Degree of Doctor in Philosophy (Ph.D.), College of Computing and Digital Media, DePaul University, 2009.

[2] S.K.Sundaram. "Requirements Tracing Using Information Retrieval", Dissertation for the Degree of Doctor in Philosophy (Ph.D.) in Engineering, University of Kentucky, 2007.

[3] S.K.Sundaram, J.H.Hayes, A.Dekhtyar, E.A.Holbrook. "Assessing traceability of software engineering artifacts". In Requirements engineering 15 (3), pp:313-335, 2010.

[4] J.H.Hayes, A.Dekhtyar, S.K.Sundaram, Advancing Candidate Link Generation for Requirements Tracing: the Study of Methods. In Software Engineering, IEEE Transactions on 32 (1), pp:4-19,2006.

[5] J.H.Hayes, A.Dekhtyar, S.K.Sundaram, S.Howard."Helping Analysts Trace Requirements: An Objective Look". In Requirements Engineering Conference. Proceedings. 12th IEEE International, 6-11 Sept. 2004 pp: 249 – 259. ISSN :1090-705X. Print ISBN:0-7695-2174-6, 2004 .

[6] S.K.Sundaram , J.H.Hayes , A.Dekhtyar. "Baselines in Requirements Tracing". In ACM SIGSOFT Software Engineering Notes 30 (4), pp:1-6, 2004.

[7] H.Sultanov. "Application Of Swarm And Reinforcement Learning Techniques To Requirements" , Dissertation for the Degree of Doctor in Philosophy (Ph.D.) in Engineering , University of Kentucky, 2013.

[8] C. Jones."Patterns of Software System Failure and Success" , Intl Thomson Computer Pr (Sd) (1 December 1995),1995.

[9] P.A.Laplante."Real-time systems design and analysis" , THIRD EDITION, A JOHN WILEY & SONS, INC., Canada,2004.

[10] J.H.Hayes, A.Dekhtyar , S.K.Sundaram, S.Howard. "On Effectiveness of User Feedback-based Information Retrieval Methods for Requirements Tracing", 2004 (TR 423-04), 2004.

[11] J.H.Hayes, A.Dekhtyar , J.Osborne ."Improving requirements tracing via information retrieval". In Requirements Engineering Conference, 2003. Proceedings. 11th IEEE International pp:138-147, 2003.

[12] H.Sultanov, J.H.Haye , W.K.Kong."Application of swarm techniques to requirements tracing" , Journal of Springer-Verlag London Limited 2011, N0:10,2011.

[13] P.Pantel , D.Lin." A Statistical Corpus-Based Term Extractor" , Advances in Artificial Intelligence, Lecture Notes in Computer Science Volume 2056, 2001, pp 36-46, 2001.

[14] M.T.Pazienza , M.Pennacchiotti, F.M.Zanzotto."Terminology extraction: an analysis of linguistic and statistical approaches" In Knowledge Mining, S.Sirmakessis (Ed.), Series: Studies in Fuzziness and Soft Computing, Vol.185, Springer Verlag, 2005.

[15] L.Ahrenberg." Term extraction: A Review Draft Version 091221", 2009.

[16] J.Tilley."A Comparison of Statistical Filtering Methods for Automatic Term Extraction for Domain Analysis" , Dissertation for the Degree of Master of Science (M.Sc.) in Computer Science, University Virginia,2008.

[17] M.F Porter." An algorithm for suffix stripping". Program 14(3):130–137,1980.

A Fast and Accurate Pupil and Iris Localization Method Usable with a Regular Camera

Gheis Mohammadi[1]
[1]Department of Computer Engineering, Science and Research branch, Islamic Azad University of Tehran
Tehran, Iran

Jamshid Shanbehzadeh[2]
[2]Department of Computer Engineering, Kharazmi University
Tehran, Iran

***Abstract*— The fundamental step in implementing an eye-based interface is the exact localization of user eye pupil and iris and separating them from other parts of the eye. This paper presents a fast adaptive iris localization scheme that can be used by low resolution cameras or webcams. This method finds the pupillary and iris boundaries accurately with low computational cost under variable and noisy lighting conditions. Based on the fact that the pixels of the pupil are darker than other regions, this method locates its boundary circle and then uses it to localize the outer boundary of the iris. The first step to identify the inner boundary is to calculate a threshold value to separate the pupil's pixels from other parts of the eye. Next, we search in the neighbourhood of the initial threshold value for the most appropriate threshold value. This value binaries image and Circular Hough Transform detects circles with certain values. After remove most circles, the most accurate one is chosen. This circle is the boundary of the pupil. Finally, the outer boundary of the iris is identified. The performance of the proposed algorithm was assessed by using it to segment both low and high resolution images. We ran the experiments on UBIRIS v1.0 also, for comparison with other reported accuracies. The experimental results show that the proposed method has a 14ms detection time and 97.76% detection accuracy with an overall accuracy of 99.025%. These results can be achieved in low resolution images also that show an improvement in pupil and iris localizing performance in comparison with well-known methods.**

Keywords: **Eye localization; Eye tracking; Image analysis; Pupil and iris boundary; Robust Adaptive localization.**

I. INTRODUCTION

Eye-based interaction is one of new areas in human-computer interaction that there has been an increased interest in its applications. The fundamental step in implementing an eye-based interface is the exact localization of user eye pupil and iris. Iris localization (IL) is commonly considered as a challenging problem and a suitable scheme should cope with several conditions such as uneven texture contrast, side effects due to the presence of eyelids and eyelashes, variable contrast between sclera and iris and between pupil and iris, some lighting points and light reflection of the iris as well as blurred boundaries of the iris.

Currently, there are a lot of related researches about iris and pupil localization.[1] Each research has different technique for extracting and localizing of iris and pupil. The most of these methods use two circles or ovals to approximate the boundaries of the iris; one for the inner (depending on the application type, it can be the pupil area or the iris area) and the other for the outer boundary (iris area or sclera region can be considered). The estimation of the location and size of these two areas is the focus of iris localization. This paper presents a fast and robust IL algorithm based on the Histogram and Hough transform. This method first detects the pupil boundary circle then localizes the outer iris boundary. To verify the performance of the proposed method, we used a well-known eye database (UBIRIS) and performed some basic experiments on it. We show in this experiments that the proposed method has acceptable accuracy and its processing time is shorter than previous methods. Then, we apply the proposed algorithm to an eye tracker, along with an low resolution camera, and we verify the effectiveness of this approach.

The organization of the rest of this paper is as follows: Section 2 introduces related works. Section 3 presents the proposed algorithm. Section 4 gives the experimental and comparative results of the proposed method and the conclusion comes in Section 5.

II. LITERATURE REVIEW

The circular pupil localization methods can be partitioned in three groups: template-based, cluster-based and Hough transform (HT) based techniques.

The template-based method usually uses complex parametric equations. In 1993, Daugman [2] presented an IR algorithm whose IL was based on an integro-differential operator to locate the pupillary and iris boundaries. The operation of his method was based on the assumption that the pupil has circular edges. Daugman‘s method is effective in high quality iris images, but its performance decreases in

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

presence of specular reflection and in case of low contrast images. Camus and Wildes [3] proposed a similar algorithm to Daugman's one. This algorithm may fail under noisy conditions or the presence of light reflections.
Some recently works on template-based pupil localization whose core is the Daugman's method[4,5] or combination of some template-based techniques are as follows: integro-differential operator by Sanchez-Avila et al.[6,7], morphological and threshold technique by Kun Yu et al[8] and Kennell et al. [9], bisection technique by Lim et al. [10], pattern matching by Emmanvel Raj.M.Chirchi et al. [11], high threshold method by Bhola Ram Meena [12], using Laplacian of Guassian (LoG) mask R. Krishnamoorthy et al. [13], threshold based Freeman's chain code by Vatsa et al. [14], active contour model by A. Ross et al [15] and A. Abhyankar et al. [16], edge based virtual circle by Boles et al. [17], wavelets based by J. Cui et al. [18] and Y. Z. Shen et al. [19], threshold based ring mask by Lye Liam et al. [20], features optimizing with PCA and SVM classifier by K N Pushpalatha et al.[21], etc. Template-based methods have high computational complexity and parameters requires reconfiguration per each database. These methods are sensitive to noise and may fails when images do not have sufficient intensity separation between eye regions.

The cluster based localization techniques typically precede two stages. First, the clustering and normalization and second, edge detection and localization of iris boundaries. Some researches whose used clustering includes texture segmentation by Kavita. [22], Hough clustering by Lili Pan et al. [23] GLCM pattern analysis technique by Bachoo et al. [24], etc. The main disadvantage of this method is that the cluster based localization requires more computation time because Preprocessing steps like clustering and normalization are time consuming.

The third category is the most common methodology, being used in several researches [25,26,27,28]. The HT based techniques first detects the edges in a region of interest and then the circular Hough transform finds the circular boundaries of the iris. The accuracy of HT based approach depend on the specific image characteristics such as brightness, and contrast. In addition, the edge-detector algorithm and its necessary tuning parameters are critical factors for localization accuracy. Improper adjusting of these parameters can lead to inappropriate boundary detection. As such, in this work, we propose a fast and robust IL algorithm based on the Hough transform that more accurately detects iris inner and outer boundaries.

III. PROPOSED METHOD

This section presents an adaptive IL algorithm that finds the pupillary and iris boundaries accurately with low computational cost under variable and noisy lighting conditions. This approach is based on the fact that the pupil's pixels (also the eyelashes and eyelids area pixels) are darker than the pixels of other regions.

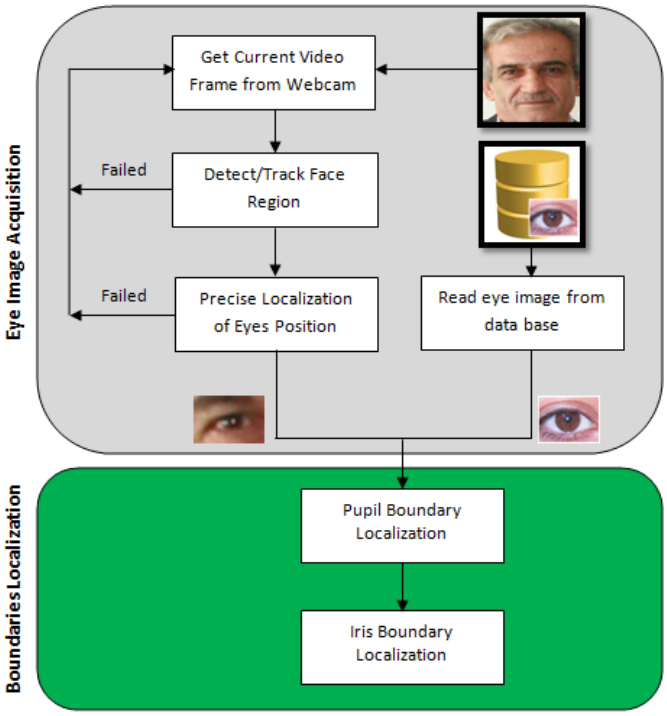


Fig. 1. The general block diagram of an iris and pupil localization system

Figure 1 shows the general block diagram of an iris and pupil localization system. The blocks in grey background prepare the eye image. The eye image can be read from an eye database or captured by a camera. This stage of the system is not on the scope of this research and our purpose is focusing on iris localization that start in next stage. The blocks in green background show our proposed method. This method first locates the boundary of the pupil then uses it to localize the outer iris boundary. Figure 2 shows the details of our method. The red blocks perform pupil localization and yellow blocks localize iris boundary.

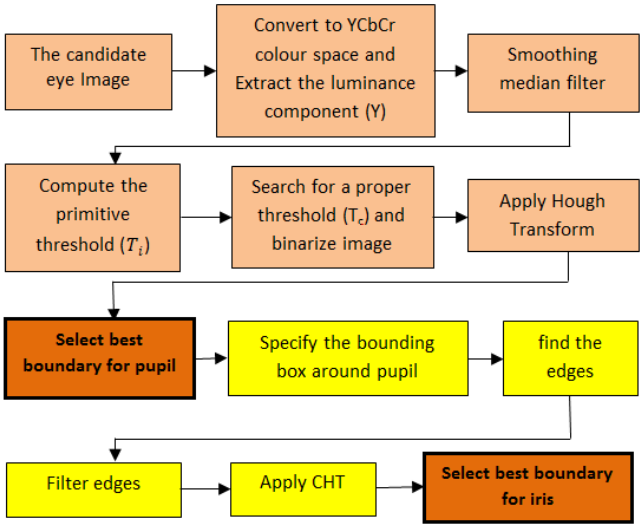


Fig. 2. The block diagram of the proposed method

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

The first block captures the candidate eye image and second block filters its Y channel (in YCbCr colour space) from the region of interest. The third block applies median filter and the next block calculates the initial threshold value. This value separates the pupil pixels from the other part of the eye image. The next block searches for the most appropriate threshold value to find the boundary of the pupil and the last red block uses HT to detects all circular boundaries. finally, the algorithm chooses best circle for pupil. The next stage is iris boundary localization start with finding the edges in a bounding box around pupil. The third yellow block filter these edges and the next block searches for circles. Finally, best circle for iris boundary are chosen.

Pupil boundary localization

The pupil is the darkest area of the eye region and its shape is a circle, so it can provide important features for eye localization. According to this feature, we can design an algorithm to find the connected dark pixels that form a circle in an eye image.

This section presents a technique based on HT to localize the boundaries of both pupil and iris. HT suffers from high computational cost. This problem is the result of a large number of search states. The new algorithm reduces the execution time by removing both the irrelevant edge points and the number of possible circles.

First, we need an eye intensity image to localize its boundaries. The luminance component (Y) of YCbCr colour space is chosen as the eye intensity image. The reasons for this selection are as follows:

1. The luminance component of YCbCr is independent of colour, so it can be employed to solve the illumination and colour variation problem.
2. In comparison with the components of the other colour spaces, the luminance component of YCbCr has a relatively higher contrast compared with skin and the white area of the eye.

The next step uses the CDF method to find a primary threshold value that separates a high percentage of eye pupil pixels from the rest of the eye pixels [28]. Next, it scans nearby primary threshold values. For each threshold value, the CHT searches for the pupil boundary. In most cases, the first detected circle is the best choice but we filter out some of the detected circles. This novel method is as follows:

1. Consider a horizontal band of the eye image, starting from one fourth to four-sixths of the height of the eye image (Fig 6.b and Fig 6.c).

2. Extract the luminance component of this band (Fig 6.d). Several definitions can be used for this transformation. This paper calculates the luminance Y as a weighted sum of RGB components. Equation (1) shows the calculation.

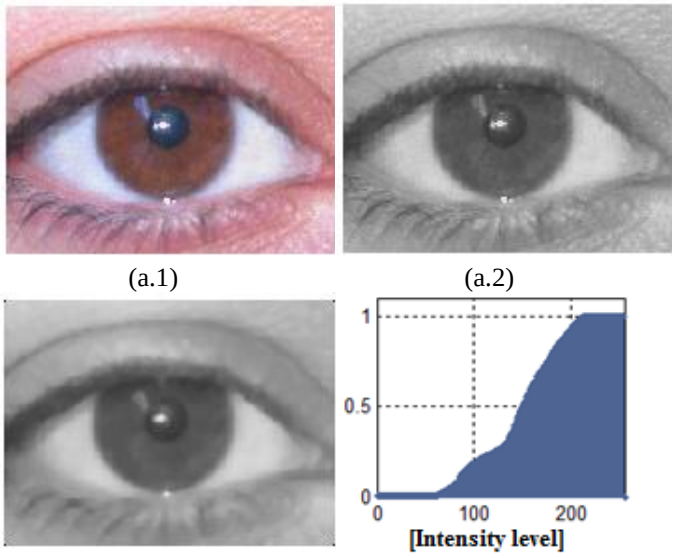
$$Y=0.299R+0.587G+0.11B \quad (1)$$

3. Apply a smoothing median filter on it. (Fig 6.e).

4. Compute the cumulative histogram 'H' of the filtered image. The cumulative histogram can be found by integrating the histogram of each of the ROI by using the following equation:

$$H(L) = \sum_{g=0}^{L} h(g) \qquad (2)$$

Where *h(g)* is the histogram representing the probability of occurrence of intensity level '*g'* and $0 < L < 255$. Figure 3 shows the cumulative histogram of an eye.



(a.1) (a.2)

(a.3) (a.4)

Fig.3. (a.1) an eye image from our database
(a.2) The luminance component (Y) of the eye image
(a.3) median filtered
(a.4) cumulative histogram of the filtered image

Compute the primitive threshold` 'T_i' where is the biggest value of *H* that is smaller than *'P'*. The parameter *'P' is a* certain initial probabilities of an eye pupil pixel and determines the probability of the pupil pixels. This means if we calculate probabilities of the all eye pixels, pupil pixels have a value around *'P'*. This parameter must be chosen between 0.02 and 0.09. In this range, the sensitivity of the overall system with respect to this parameter is low [28]. Figure 4 shows an eye image after thresholding and its corresponding *'P'* value.

As we can see in fig.4, using parameter 'P' the pupil pixels remains in eye image and other pixels are removed.

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

Despite the fact that this parameter is a great tool for extracting pupil pixels, but can't remove the pixels of other area permanently and output pixels contains undesired parts of eyebrow or other dark areas of eye. Therefore, we must do some extra steps to find pupil boundary.

In step 6 to 9, we uses a loop and search around initial threshold 'P' (from T1 to T2) to find a binarized image which eye pupil boundary are better extracted on it.

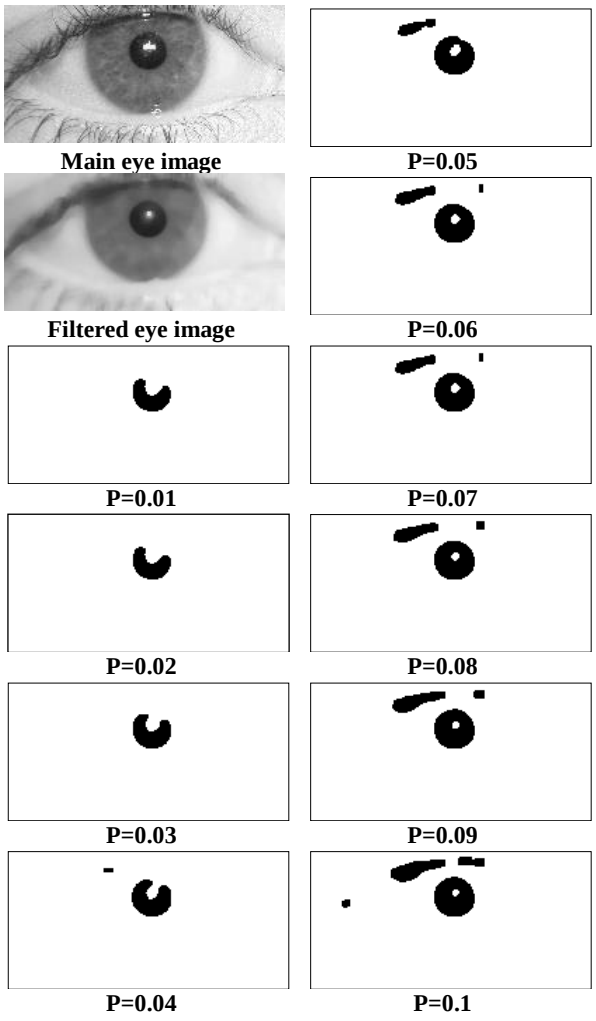


Fig.4. threshold images of an eye image with respect to different *'P'* values.

5. Consider the threshold 'T_1' where $T_1 = \alpha T_i$, and 'T_2' where $T_2 = \beta T_i$, and where $0.5 < \alpha < 1$ and $1.1 < \beta < 1.5$. There exists a nonlinear relation between these two parameters α *and* β and the computational cost and accuracy of our localization algorithm. The experimental results show this nonlinearity. We start the algorithm by choosing $\alpha = 0.9$ and $\beta = 1.1$.

6. We set T_c and the *'stw'* to T_1 and 1 respectively. T_c is the current threshold for binarizing image and can change from T1 to T2. *'stw'* Parameter is the amount by which T_c is incremented each step.

7. The filtered image is binarized according to T_c and the result undergoes CHT to find circles with a radius smaller than 25% of the height of the filtered eye image. The radius of eye pupil circle is less than 25% of its height. Therefore, in this step the algorithm can localize pupil boundary.

8. If the CHT method detects any circles then $T_a=T_c$ (see Fig 6.f for binarizing the image using T_a) and we go to step 10, otherwise the following step should be performed.
   $T_c = T_c +$ *'stw'* (3)
   If $T_c \leq T_2$ then we go to step 8 (search again), otherwise there is no eye in the image or the eye is closed and we finish the algorithm.

9. If the previous steps detect one circle, then that is the boundary of the pupil. It's because of this fact that pupil pixels are the most dark pixels in eye image and by binarizing the image using low thresholds, other pixels are removed and remained area is a part of pupil. In case of detecting more than one circle, we filter out some circles and choose one of them.

We can use the following solution in filtering and choosing the best circle.

- The biggest object in the binarized image includes part of the pupil. We choose a circle that best covers this region. This circle contains more black pixels from the last thresholded eye image. Applying some geometrical constraints can be useful.

Figure 5 shows the pseudo code of the pupil localization. The last two functions, *'filterOutBadCircles'* for removing inaccurate circles and *'selectBestCircle'* for choosing the most accurate circle, performs step 10. Figure 6 shows the results of all steps of the algorithm.

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

```
ALGORITHM localizePupil()
{
//set initial values
SET P to 0.05
SET alpha to 0.9
SET beta to 1.1
SET stw to 1
SET found to false
//start localization
mainEyeImage ← getEyeImage()
Ycomponent ← extractYComponent(mainEyeImage)
filteredImage ← smoothMedianFilter(Ycomponent)
cHist←computeCumulativeHistogram(filteredImage)
Ti ← computePrimitiveThreshold(cHist,P)
T1 ← alpha * Ti
T2 ← beta * Ti
FOR Tc = T1 to T2 with step stw
img ← binarizeImage(filteredImage,Tc)
edgeDetectedImage ← cannyEdgeDetect(img)
circles ← applyCHT(edgeDetectedImage)
SET n to number of detected circles
IF n>0 THEN
SET found to true
IF n=1 THEN
RETURN circles[0]
Else
filterOutBadCircles(img,circles)
C ← selectBestCircle(img,circles)
RETURN C
END IF
END IF
END FOR
IF found=false THEN
PRINT "no eye or eye is closed"
RETURN null
END IF
}
FUNCTION computePrimitiveThreshold(cumHist,P)
{
FOR L=1 to 255
IF cumHist(L)>P THEN
RETURN L-1
END IF
END FOR
}
FUNCTION filterOutBadCircles (binarizedEyeImg ,
circles)
{
C ← the center of biggest object in binarizedEyeImg
N ← count of circles
maxD = 0.1 * binarizedEyeImg.Width
FOR I = 0 to N-1
IF DISTANCE(circles(I).center,C) > maxD THEN
removeCircle(I)
END IF
END FOR
}
FUNCTION selectBestCircle (binarizedEyeImg , circles)
{
N ← count of circles
SET B to 0
maxBlackPixels ← 0
bestCircle ← null
FOR I = 0 to N-1
B ← the count of black pixels in circles(I)
IF B > maxBlackPixels THEN
maxBlackPixels ← B
bestCircle ← circles(I)
END IF
END FOR
RETURN bestCircle
}
```

Fig.5. Our pupil localization algorithm pseudo code

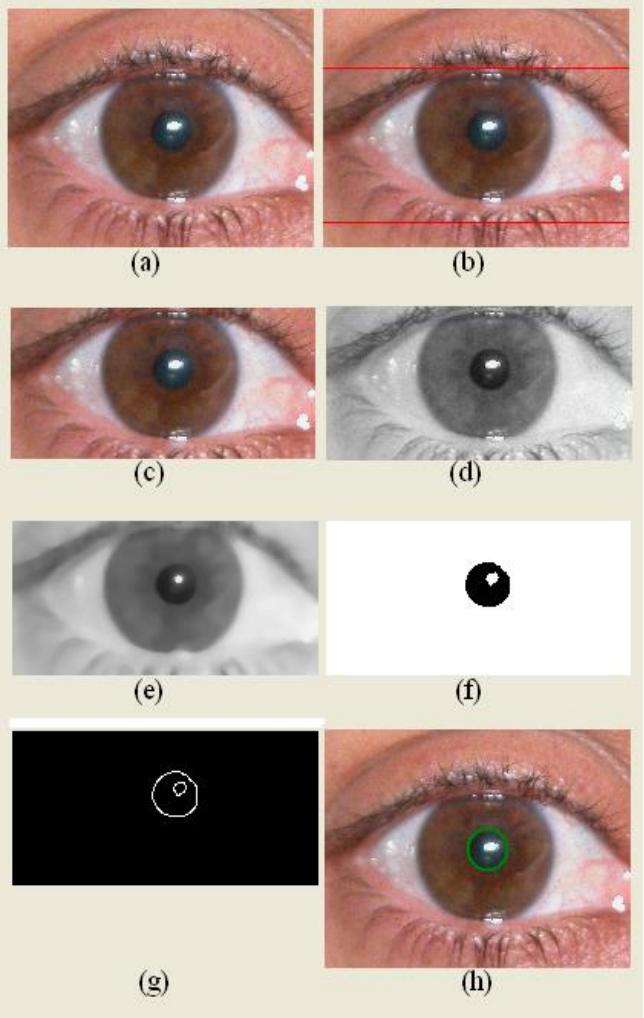


Fig.6. Step-by-step result of the proposed pupil localization method.
a) The candidate eye image
b) Considered horizontal band of the eye image
c) Cropped band
d) Extracted Y channel
e) Applied smoothing median filter
f) Thresholded image
g) Edge detected
h) Pupil localized eye image

Iris boundary localization

This step assumes that the previous steps have detected the inner circle of iris (pupil area) and it considers an area with the centre of the pupil circle to localize the iris. Within this area the edge points are detected and then the CHT searches for a circle with a certain radius size. The iris outer boundary detection is performed by using the following steps:

Step 1: We define the bounding box with the centre of the pupil area and four times its size.

Step 2: Find the edge points within this box by using the Canny edge detector.

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

Step 3: Filter out some of these edge components. (Fig 7.c)

Step 4: Apply CHT and then search for a circle whose center is within the pupil area and whose radius 'R' is $2r_{pupil} \leq R \leq 4r_{pupil}$ (see Fig.7.b) where r_{pupil} is the radius of the pupil circle. This step returns the iris circle (Fig.7.d). If no circle is detected we can increase R and search again.

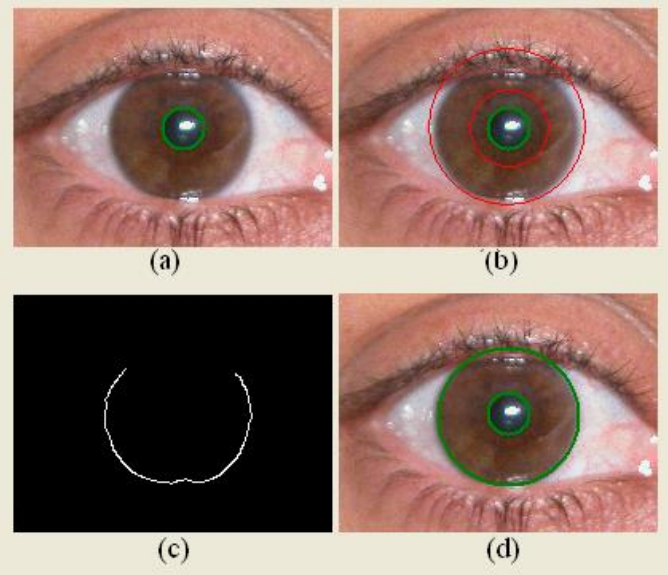


Fig.7. Step-by-step result of the proposed iris localization method.
a) The candidate's eye image with pupil detected
b) Considered bands of the iris
c) Threshold image and some filtered edge components
d) Iris localized eye image

IV. EXPERIMENTAL RESULTS

The algorithm is developed using the visual C#.net 2010 programming language and the OpenCV computer vision library wrapper, 'EMGUCV'. It is tested on the Centrino dual core 2.0 GHz CPU with windows XP service pack 3 and 2 GB DDRII Ram. The 'UBIRIS V1' [29] iris database has been selected for experiments. This database is famous and consists of different eye characteristics such as environmental light changes, noise and, focusing problem. This database consists of two parts: Part 1 with 1214 images and Part 2 with 663 images.

Tables 1 and 2 show the variations of detection time and accuracy of our method (for 500 random eye images from UBIRIS database) with respect to different values of $'\alpha' and '\beta'$ and '*stw*'=1 and '$P'=0.05$. Figures 8 and 9 show the results graphically.

We can see that tuning these values can decrease the detection time by more than 90 ms and increase the detection accuracy by about 2%. Therefore, we choose reasonable values for these parameters. In our experiments, these values are α=0.9 and β=1.1. These can be changed by varying *'P'*.

Table I. The effects of two parameters, *'α'* and *'β'*, on the search time(ms) of algorithm.

		β				
		1.1	**1.2**	**1.3**	**1.4**	**1.5**
α	**0.5**	39.30	42.58	44.56	45.28	47.70
	0.6	27.97	30.80	32.09	33.76	35.54
	0.7	21.10	22.88	25.28	26.86	27.70
	0.8	16.66	18.37	19.54	21.11	21.73
	0.9	14.93	17.67	19.92	21.64	22.78

Table II. The accuracy (%) of the algorithm based on the value of two parameters *'α'* and *'β'*.

		β				
		1.1	**1.2**	**1.3**	**1.4**	**1.5**
α	**0.5**	95.8	96.0	95.6	95.2	94.6
	0.6	96.6	96.6	96.0	95.8	95.8
	0.7	96.2	96.4	96.2	96.0	96.2
	0.8	96.8	96.6	96.4	96.6	96.4
	0.9	96.8	96.2	96.4	96.0	96.2

We tested our method on the UBIRIS database. Table 3 shows the results of this experiment. In this experiment, we assume a localization error of less than 7 pixels as correctly localized.

Table III. THE RESULTS OF OUR EXPERIMENT AGAINST UBIRIS.V1

α =0.9 and β =1.1 and stw=1		
Localization Properties	**Session 1**	**Session 2**
Correctly Localized	1188	571
Failed	26	92
Min Localization Time	13 ms	14 ms
Max Localization Time	46 ms	61 ms
Mean Localization Time	14.21 ms	17.19 ms
Accuracy	97.86%	86.12%
α =0.9 and β =1.1 and stw=2		
Correctly Localized	1162	561
Failed	52	102
Min Localization Time	5 ms	7 ms
Max Localization Time	18 ms	23 ms
Mean Localization Time	7.34 ms	8.01 ms
Accuracy	95.71%	84.61%

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

As we can see in Table 3, with a small step width, the overall accuracy of our method in session1 is 97.86% and in session2 is 86.12% and the proposed method successfully localized 1188 images out of 1214 in session1 and 571 images out of 663 in session 2. The worst result in session2 is due to the very high light variations and poor intensity separability between sclera, iris and pupil in the images of session 2. In this situation, the localization time of our algorithm is less than 16 ms for α=0.9 and β=1.1. This time is appropriate for online processing with a frame rate of more than 60 frames per second. By increasing the value of the '*stw*' parameter to 2, the localization time improves around twice and the accuracy of the method decreases about 2 percent.

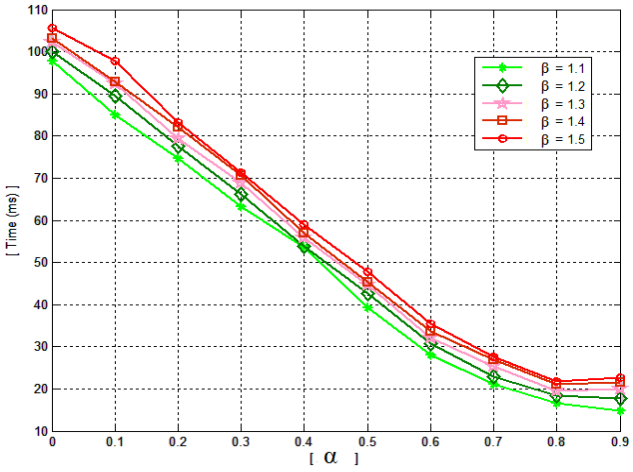


Fig.8. The detection time (ms) of our localization algorithm with respect to variations of $'\alpha'$ and $'\beta'$.

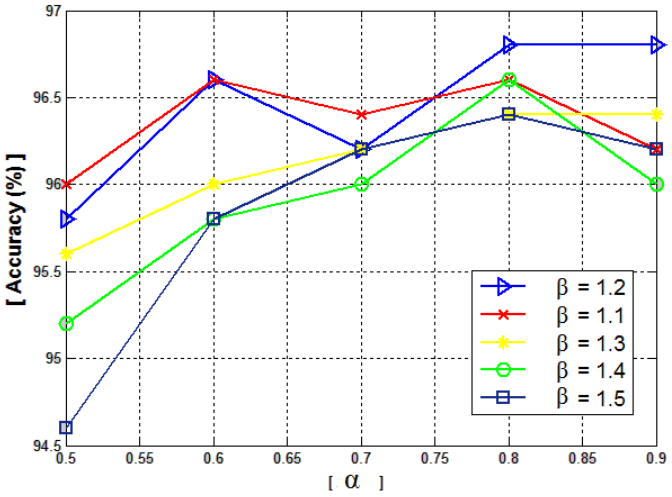


Fig.9. The accuracy(%) of our localization algorithm with respect to variations of $'\alpha'$ and $'\beta'$.

The next experiment describes a localization error to show the accuracy of the algorithm by Equation (4).

$$detection\ error = \frac{\max\{|R - R'|, \|C - C'\|\}}{W} * 100 \quad (4)$$

In Equation (10), R and C are the actual radius and centre of the pupil boundary. '*R'* and '*C'* are also the radius and centre of the localized area for the proposed algorithm and *W* is the width of the given eye in pixels. We define the efficiency of the algorithm for a given detection error as the number of eye images less than the detection error of our method. Table 4 shows the efficiency of the proposed algorithm against the two sessions of the UBIRIS database. This experiment assumes that *'W'* is equal to the width of the given eye image. The width of all images in the UBIRIS database is 200 pixels.

Table IV. THE EFFICIENCY OF THE PROPOSED ALGORITHM

Detection Error	Session 1 (1188 true localized)		Session 2 (571 true localized)	
	count	percentage	count	Percentage
0.0%	408	34.34%	327	57.27%
0.5%	359	30.22%	89	15.59%
1.0%	207	17.42%	66	11.56%
1.5%	146	12.29%	38	6.65%
2.0%	36	3.03%	19	3.33%
2.5%	19	1.6%	23	4.03%
3.0%	3	0.25%	6	1.05%
3.5%	10	0.84%	3	0.53%
4.0%	0	0%	0	0%

The results show in low detection errors the efficiency of algorithm is good and the overall efficiency is high. Figures 10 and 11 show respectively results of inaccurate iris localization in session 1, and session 2.

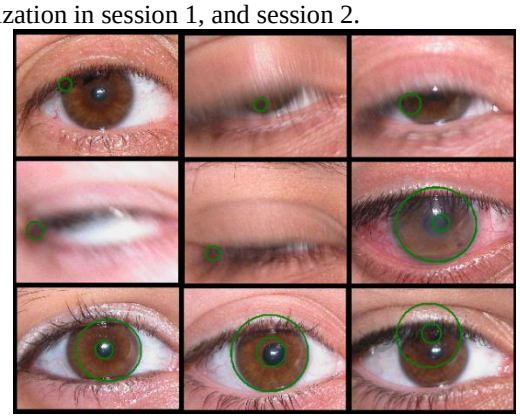


Fig.10. inaccurate localization in session 1

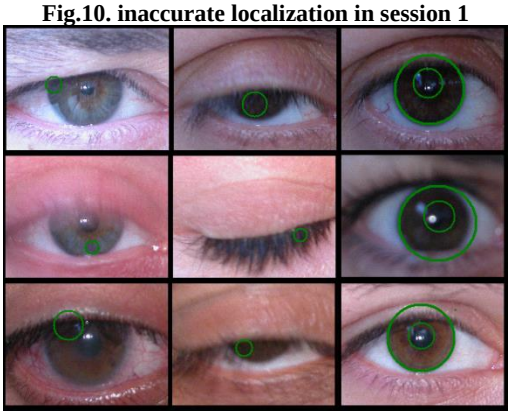


Fig.11. inaccurate localization in session 2

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

Table V. THE ACCURACY RATE AND THE EXECUTION TIME OF THE PROPOSED METHOD

Method	Parameters	Session 1	Session 2	Time for session1 (ms)	Time for session2 (ms)	Implementation Notes
[3]	Hysteresis Thresholds: Hi=50, Low=44, Gaussian Kernel Dimension=5	86.64%	73.26%	3600	3600	• Cases of segmentation failure are determined visually
[30]	Constant parameters	98.13%	87.48%	650	650	• 200 iris images selected from Session 1 are used for training • the remaining 1014 images are used for testing. • the parameters learned are applied directly to Session 2 without any adjustment.
[31]	Constant parameters	98.43%	85.82%	800	800	• the accuracy of the localization is determined visually.
[32]	4x downsized ratio	95.46%	87.03%	2970	2970	• The assessment of the algorithm was done by visually inspecting • The images have been reduced in size by a factor of 4 to increase the speed of the algorithm. • After some modifications, the accuracy obtained for session 1 is 92.36% and for session2 is 83.96%, with an average segmentation time of only 0.48 seconds.
ours	α=0.9 β=1.1 stw=1	97.87%	86.11%	14.21 [13-46]	17.19 [14-61]	• Are fully reported in the experimental result section

Next experiment compares the performance of the proposed algorithm with Wildes [3] (a classic HT based method) and three other method introduced in [30][31],[32]. Table 5 shows the accuracy rate and the execution time of the proposed method. The best rating in every item is marked by green color. The accuracy rate of our method is much higher than Wildes' method but, it is less than [30], [31] in session1 and [30],[32] in session2. However, in terms of speed, the proposed method is much faster. This is because of the different proposed steps that applied to reduce the searching space in CHT method. The reported performance of Wildes' method was extracted from [30].

The next experiment tests this method in an online eye tracking system by using a Microsoft LifeCam VX700 webcam for capturing frames. Each frame of the input video is processed by three different AdaBoost classifiers to detect the locations of the user's face and their left and right eye. We use a set of simple geometrical constraints to verify the detected regions. Sometimes the user turns his/her face without placing it in the frontal position. In these cases, the classifier fails to detect the face location and, we use the Camshift algorithm [23] to track the face, which gives us the rotation angle of the head. Our method supposes that the eyes are located in the second quarter of the head and we use the head angle to approximate their location. Table 6 shows the average execution time of each part for a single frame. The pupil localization part is very fast and its execution time is about 13 milliseconds.

Table VI. THE AVERAGE EXECUTION TIME OF EACH PART

Algorithm Parts	Time(ms)
Face detection (Using Haar Cascade Classifier)	27.91
Face tracking (Using Camshift)	14.47
Left eye detection (Using Haar Cascade Classifier)	30.20
Left Eye tracking (Using Haar Cascade Classifier and some simple geometrical constraints)	16.83
pupil boundary detection	**13.2**
total average execution (Face & Eye Tracking + Eye pupil localization)	44.5

Another experiment was performed with a total of 10 participants (8 males and 2 females) with a wide age range and some of them wearing glasses. The participants sat approximately 60 cm from the camera. A black screen with a moving red point was shown to them (Figure 12). The red point was randomly moved on the screen and we asked the users to keep their heads steady and gaze and track this point. Our application captured 9000 frames (5 minutes with 30 frames per seconds) for every person and processed these frames in real time. Table 7 shows the results.

The detection accuracy and the overall detection accuracy is calculated using (5) and (6), respectively.

$$Detection\ Accuracy = \frac{TP}{TP+FN} * 100\% \quad (5)$$

$$Overall\ Accuracy = \frac{TP+TN}{TP+FP+FN+TN} * 100\% \quad (6)$$

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

TABLE VII. **THE DETECTION ACCURACY AND THE OVERALL DETECTION ACCURACY OF OUR METHOD**

Subject ID	Left Eye Was Opened			Left Eye Was Closed			Detection Accuracy	Overall Accuracy
	Total Frames	True Boundary Localized (TP)	Wrong Boundary for Pupil was Localized (FN)	Total Frames	Wrong Boundary is Detected as Pupil (FP)	Reported as "No Pupil" (TN)		
1	8683	8546	137	317	220	97	98.42%	96.03%
2	8615	8526	89	385	184	201	98.97%	96.97%
3	8256	8166	90	744	180	564	98.91%	97.00%
4	8511	8434	77	489	103	386	99.10%	98.00%
5	8727	8689	38	273	54	219	99.56%	98.98%
6	8495	8317	178	505	131	374	97.90%	96.57%
7	8686	8573	113	314	116	198	98.70%	97.46%
8	8508	8472	36	492	41	451	99.58%	99.14%
9	8414	8398	16	586	57	529	99.81%	99.19%
10	8470	8412	58	530	97	433	99.32%	98.28%
Total	**85365**	**84533**	**832**	**4635**	**1183**	**3452**	**99.025%**	**97.76%**

TP: Number of frames with correctly localized eye pupil boundary (true positive)
FN: Number of frames with open eye which the method did not localize (false negative)
FP: Number of frames with closed eye that are a circle and are localized as pupil (false positive)
TN: Number of frames with closed eye that are correctly reported as no pupil (true negative)

Fig.12. A participant sat in the front of camera and a black screen with a moving red point was shown to him

Because, the iris capture devices are mostly exposed to natural illumination or other variant circumstances, sometimes these conditions affect the quality of the iris images and further impact on the localization result. We used this method in three lighting conditions and Figure 13 presents the experimental results. It shows that variant lighting conditions do not affect the overall performance and accuracy of the algorithm.

The experimental results show that the proposed method has a 97.76% detection accuracy with an overall accuracy of 99.025%. These experimental results are promising and show an improvement in pupil and iris localizing accuracy and execution time in comparison with similar well known methods.

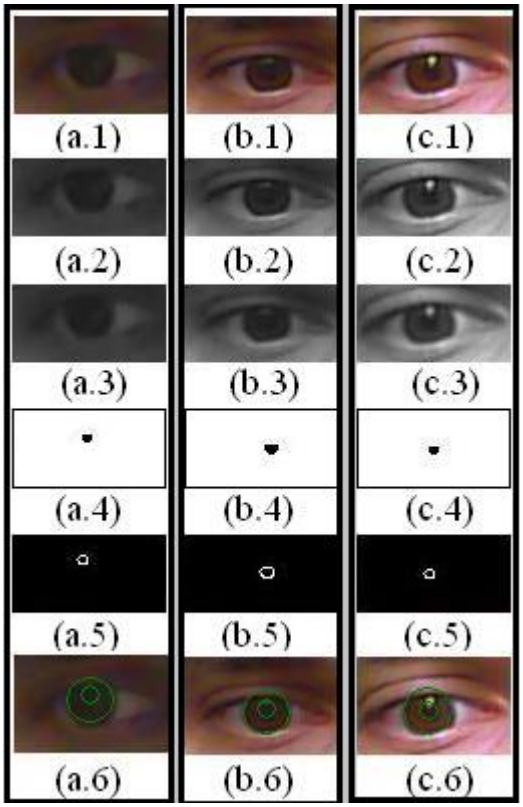


Fig.13. step by step results of proposed method used in eye tracking system for three light conditions. (a):Dark condition (b):Median lighting condition (c):Bright condition
(a.1, b.1, c.1) The user eye image
(a.2, b.2, c.2) Cropped horizontal band of the Y component
(a.3, b.3, c.3) Applied smoothing median filter
(a.4, b.4, c.4) Threshold image
(a.5, b.5, c.5) Edge detected
(a.6, b.6, c.6) Pupil and Iris localized eye image

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

From the experiments, some situations occasionally decrease the accuracy of our algorithm. The first is the result of a swift movement of the head. In this situation, our algorithm is unable to correctly localize the eye. This is because the eye images are blurry so that the skin colours blend with the colours of the eye areas. The other situation is when the user bows his/her head or changes the focus to the lower area (with respect to the camera position) so that the eyelids are captured partially close. Other reasons for the failure of our method are high reflections, high variations of luminosity and poor focus.

V. CONCLUSIONS

This paper proposed a fast iris localization method based on CHT. The proposed method consisted of these steps: first, a horizontal part of the eye image was considered and its luminance version (Y component) was extracted. Second, a smoothing filter was applied on it and a primitive threshold was obtained. Third, in the neighbourhood of this threshold value, a certain threshold value for localizing the pupil boundary was considered and, finally, by the use of a localized pupil circle, the iris region was detected. The proposed approach improved both localization time and accuracy in comparison with well-known methods.

REFRENCES

[1] Ruggero Donida Labati, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti, Iris segmentation: state of the art and innovative methods, In Cross Disciplinary Biometric Systems, C..Liu, and V.K Mago (eds.), Springer-Verlag, 2012.

[2] J. Daugman, "High Confidence Visual Recognition of persons by a test of statistical independence", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, pp.1148-1161, 1993.

[3] Camus, T.A. and R. Wildes, Reliable and fast eye finding in closeup images. IEEE 16th Int. Conf. on Pattern Recognition, Quebec, Canada, 2004: p. 389-394.

[4] J. Daugman, "Demodulation by complex-valued wavelets for stochastic pattern recognition," International Journal on Wavelets, Multiresolution and Information Processing,Vol. 1, No.1, pp. 1-17, 2003.

[5] J. Daugman, "How iris recognition works," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No.1, pp. 21-30, 2004.

[6] C. Sanchez-Avila, and R. Sanchez-Reillo, "Multiscale Analysis for Iris Biometrics", Proceedings of 36th International Conference on Security Technology, Vol. 1, pp. 35-38, 2002.

[7] Sanchez-Avila, and Sanchez-Reillo R, "Two different approaches for iris recognition using Gabor filters and multiscale zero-crossing representation", Pattern Recognition, Vol. 38, pp. 231-240, 2005.

[8] Kun Yu, Zeyu Xu and Lixin Xu, An Iris Localization Algorithm based on Morphological Processing, International Journal of Security and Its Applications Vol. 9, No. 3 (2015), pp. 153-162 , 2015

[9] L. R. Kennell, R. W. Ives, and R. M. Gaunt, Binary morphology and local statistics applied to iris segmentation for recognition, in Proc. Of 13th Annual International Conference on Image Processing. IEEE,2006.

[10] S. Lim, K. Lee, O. Byeon, and T. Kim, "Efficient Iris Recognition through Improvement of Feature Vector and Classifier", Journal of Electronics and Telecommunication Research Institute, Vol. 23, No. 2, pp. 61 – 70, 2001.

[11] Emmanvel Raj.M.Chirchi, Dr.R.D.Kharadkar, Biometric Iris Recognition for Person Identification using Cumulative Sum Algorithm, International Journal of Scientific & Engineering Research Volume 3, Issue 5, May-2012

[12] Bhola Ram Meena, "Personal Identification based on Iris Patterns", Ph.D Thesis, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, 2004.

[13]R. Krishnamoorthy and D. Indradevi, A New Snake Model for Pupil Localization Using Orthogonal Polynomials Transform, International Journal of Computer Theory and Engineering, Vol. 5, No. 1, February 2013.

[14] M. Vatsa, R. Singh, and A. Noore, "Reducing the False Rejection Rate of Iris Recognition using Textural and Topological Features", International Journal of Signal Processing, Vol. 2, No. 2, pp. 66-72, 2005.

[15] A. Ross and S. Shah, Segmenting non-ideal irises using geodesic active contours, in Special Session on Research at the Biometric Consortium Conference, 2006, Biometric Symposium, 2006.

[16] A. Abhyankar and S. Schuckers, Active shape model for effective iris segmentation, in Proceeding of the SPIE, vol. 6202, 2006.

[17] W. W. Boles and B. Boashash, "A Human Identification Technique using Images of the Iris and Wavelet Transform", IEEE Transactions on Signal Processing, Vol. 46, No. 4, pp. 1185-1188, 1998.

[18] J. Cui, Y. Wang, T. Tan, L. Ma, and Z. Sun, A fast and robust iris localization method based on texture segmentation, in SPIE Defense and Security Symposium, vol. 5404, pp. 401–408 977, 2004.

[19] Y. Z. Shen, M. J. Zhang, W. J. Yue, and M. H. Ye, A new iris locating algorithm, in Proc. of Intl. Conf. on Artificial Reality and Telexistence–Workshops, vol. LNCS 3617. Springer, 2006, pp. 438–441.

[20] Lye Liam, Ali Chekima, Liau Fan, and Jamal Dargham, "Iris recognition using self-organizing neural network, Proceedings of IEEE Student Conference on Research and Developing Systems', pp.169-172, 2002.

[21] K N Pushpalatha, Aravind Kumar Gautham, D.R.Shashikumar, K.B ShivaKumar, Iris Recognition System with Frequency Domain Features optimized with PCA and SVM Classifier, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, September 2012

[22] Kavita Khobragade , A New Technique for Fast and Accurate Iris Localization, International Journal of Innovations in Engineering and Technology (IJIET), Volume 3 Issue 3 February 2014

[23] Lili Pan and Mei Xie, "Iris localization through Hough clustering", International Journal of Physical Sciences, Vol. 8(19), pp. 984-988, 23 May, 2013

[24] Asheer Kasar Bachoo, and Jules-Raymond Tapamo, "Texture detection for segmentation of iris images", Proceedings of Annual Research Conference of the south African Institute of Computer Information Technologists, pp. 236-243, 2005.

[25] Shamsi, M., P.B. Saad, and A. Rasouli, Iris Image Localization using Binning Approach. Proceedings of 2008 Student Conference on Research and Development (SCOReD 2008), 2008.

[26] Gajanan Pandurang Khetri, Dr.Vrushsen P. Pawar, Dinesh Chandra Jain, Human Computer Interface Through Biometric Iris Recognition System, International Journal of Computer Science & Engineering Technology (IJCSET), 2012

[27] Farmanullah Jan, Imran Usman, Shahrukh Agha, Iris localization in frontal eye images for less constrained iris recognition systems, Digital Signal Processing Journal, Volume 22, Issue 6, Pages 971–986, December 2012.

[28] G. Sathish, Dr. S. Narmadha, Dr. S.V.Saravanan, Dr. S. Uma Maheswari, Multi-algorithmic IRIS Recognition, International Journal of Computer Applications (0975 – 8887) Volume 38– No.11, January 2012.

[29] H. Proena and L. Alexandre, UBIRIS: A noisy iris image database, in 13th International Conference on Image Analysis and Processing - ICIAP 2005, vol. LNCS 3617. Springer, 2005, pp. 970–977.

[30] P. Li and X. Liu, An incremental method for accurate iris segmentation, in Proc. of 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[31] R. Meenakshi Sundaram , Bibhas Chandra Dhara , Bhabatosh Chanda , A fast method for iris localization, Second International Conference on Emerging Applications of Information Technology, 2011

[32] P. Radu, K. Sirlantzis, G. Howells, S. Hoque, F. Deravi, AVersatile Iris Segmentation Algorithm, In proceeding of: BIOSIG 2011 - Proceedings of the Special Interest Group on Biometrics and Electronic Signatures, 08.-09. September 2011.

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

Preserving Trajectory Privacy in Participatory Sensing Applications

Gauri R Virkar
Department Of Computer Engineering
BSIOER, Wagholi
Pune, India

Sanchika A Bajpai
Department Of Computer Engineering
BSIOER, Wagholi
Pune, India

Abstract— **With the advancement of technology in fields of wireless communication, different mobile communicating devices equipped with variety of embedded sensors and powerful sensing have been emerged. Participatory sensing is the process that enables individuals to collect, analyze and share local knowledge with their own mobile devices. Although the use of participatory sensing offers numerous benefits on deployment costs, availability, spatial- temporal coverage, energy consumption and so forth, it has certain threats which may be compromise the participator's location and their trajectory data. Henceforth, to ensure the participators' privacy is the most urgent task. The existing proposals emphasized more on participators' location privacy and very few of them consider the privacy of the trajectories. The theoretical mix zones model are been improved by considering time factor from the viewpoint of the graph theory and mix zone graph model has been presented. This model considers only sensitive trajectories for providing privacy thereby reducing overall information loss and storage space. Further, instead of defining single mix zone graph model, multiple mix zones are created in order to enhance the privacy of the participator's trajectories.**

Keywords- **Location privacy, mix zone graph model, multiple mix zone, participatory sensing, trajectories**.

I. INTRODUCTION

THE growth of mobile phones along with their pervasive connectivity leads to the development of a new sensing technology model called as participatory sensing[1] systems. Here mobile devices carried by the individual's acts as a sensor thereby eliminating the need of deploying sensors at particular areas. Participatory Sensing facilitates the participator to sense, analyze, collect and share the sensed information from their surrounding environment using their mobile phones. For example mobile phones may report actual (continuously) temperature or sound level; likewise, vehicles may notify about traffic conditions.

The vast amount of trajectory data gets collected and progressively increases as the participators sense the data. Trajectories are defined as the path followed by the moving object which is generally represented by (x, y, t) where x and y are the location coordinates and t denotes the timestamp. In typical participatory sensing applications, the data reports generated as an output may reveal participators' spatial temporal information. Adversary can obtain some valuable results from the published trajectories. The collected data may be used to deduce private information about the user. So to ensure the participators' privacy is the most urgent task. The gathered information is very crucial to the participatory sensing systems as their deficiency endangers the success of such systems. Therefore the need is to preserve the privacy of the participatory sensing users by protecting their trajectories.

Mix Zone Graph Model [2] is one of the existing approach for providing privacy to the trajectories of the participators. A mix zone is a region where no applications can track user movements. It is the region where the users can change their pseudonyms without being observed by the adversaries. A pseudonym is a uniquely generated random number. Each participator enters a mix zone with a pseudonym and exits the mix zone with other pseudonym. The use of pseudonym breaks the continuity of a user's location exposure thereby protecting the future locations of the users. However, existing mix zone model solutions mainly focus on the development of single mix zone. Henceforth, for providing more security multiple locations are selected for applying mix zone graph model. Thus, multiple mix zone model [3] is used for providing maximal privacy to the trajectories of the participatory sensing users.

A. Location Privacy

Location privacy is defined as the ability to prevent other unauthorized parties from learning one's current or past location. Traditionally, privacy of personal location information has not been a critical issue but, with the advancement of location tracking systems capable of following user movement twenty-four hours a day and seven days a week, location privacy becomes crucial: records of everything from the particular rack a person visit in the library to the clinics a person visit in a hospital can represent a very invasive list of data. Numerous systems could figure out the location of

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

person. One of several original systems designed for position following could be the Global Positioning System (GPS). This technique makes use of satellites to aid devices figure out their own position. Generally, automated digital devices obtain information either through communication, observation, or inference.

B. Trajectory Privacy

A trajectory is the path that a moving object follows through space as a function of time. Example of trajectories could be monitoring of wild animals, birds, people, a soccer player, etc. Trajectories may be uni-dimensional or perhaps multi-dimensional. Participatory sensing systems primarily depend on the collection of information across large geographic areas. The sensor data uploaded by participators are usually tagged with the spatial-temporal information when the readings were recorded the published trajectories for decision making. For example, merchants may possibly decide where to build a food store that could produce maximum gain by analyzing trajectories associated with consumers in a selected spot and also the Department of Transportation can make an optimized vehicle scheduling strategy by monitoring the trajectories connected with motor vehicles. However, it will add considerable threats to the participators' privacy. Adversary may perhaps examine the particular trajectories which contain abundant spatial-temporal background information to be able to link numerous reports that are collected. Hence, it is crucial to be able to unlink the particular participators' identities from sensitive data collection locations.

C. Existing Technique Limitation

TrPF, Trajectory Privacy Preserving Framework for Participatory Sensing Applications, is an existing approach which preserves the trajectories of the participators by applying Mix zone Graph Model at a single sensitive location. The problem here is that if an adversary is successful to guess the pseudonym of this single location Mix Zone Graph Model, the whole trajectory can be inferred.

D. Our Observation

Instead of applying Mix Zone Graph Model at single sensitive location, multiple locations can be considered as the candidates for applying Mix Zone Graph Model. As the number of locations increases, the number of pseudonyms to be cracked by an adversary increases. Thus the probability of successful attack by an adversary is reduced. An attack is said to be successful if an adversary is able to crack all the pseudonyms used in the corresponding mix zones. Consider a scenario where Mix Zone Graph Model is applied at three locations. The adversary will be able to deduce the whole trajectory only when he/she will be able to crack the pseudonyms at all three locations. Hence, as the number of mix zones increases the number of pseudonyms to be identified increases eventually increasing the privacy level.

E. Our Solution

In this paper, we propose an approach for preserving trajectories of the participators by applying Mix Zone Graph Model at multiple locations thereby enhancing the privacy level of the participators. Further, due to cost constraints, not all point of interests can be considered as the candidates for applying Mix Zone Graph Model. So the solution for selection and placement problem of the number of mix zones to be considered is been addressed here.

F. Our Contribution

Our contribution in this paper is as follows:-

- To secure location and trajectory privacy of the participatory sensing user by applying Mix Zone Graph Model.
- To secure multiple sensitive locations of the participatory sensing user.
- To prove that privacy of the user can be enhanced by protecting multiple sensitive locations instead of single sensitive location.

The remainder of this paper is organized as follows. Section II discusses about the related work. In Section III, implementation details are provided. Section IV discusses about the result work. Finally, the paper is concluded and future work is been given in Section V.

II. RELATED WORK

In the literature there exist several approaches to protect the particular position of the user. Some of them are discussed below-

A. Location Privacy Protection

There are several works that analyze the location privacy preserving schemes. They can be classified into the following aspects.

1) Obfuscation: It is defined as the means of intentionally degrading the quality of information about an individual's location in order to protect that individual's location[4].

2) Mix Networks: Mix Networks [4] uses anonymizing channels to de-link reports submitted by sensors before they reach the applications. In other words, Mix Networks act as proxies to forward user reports only when some system-defined criteria are met. Mix Network may wait to receive k reports before forwarding them to the application, e.g., to guarantee k-anonymity. However, the anonymity level directly depends on the number of reports received and "mixed" by the Mix Network. They rely on statistical methods to protect privacy and do not guarantee provably-secure privacy. In addition, there could be situations where a moderately long

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

time could pass before the desired level of anonymity is arrived at (when "enough" reports have been gathered). Accordingly, Mix Networks might strikingly diminish framework throughput and can't be utilized as a part of settings where regular reports are needed.

3) K-Anonymity: k-anonymity is a wide-spread general privacy concept not limited to location privacy. It gives the assurance that in a set of k objects (in this case, mobile user) the target object is indistinguishable from the other k – 1 object. Subsequently, the likelihood to distinguish the target user is 1/k. The thought behind k-anonymity is that a user reports a obfuscation region to a customer containing his position and the positions of k – 1 different customers rather than his exact position that is secured by a pseudonym. As an example consider that Alice is currently at home and queries a location based service for the nearest cardiology facility. Without utilizing anonymization, this inquiry could reveal to the customer implementing the service that Alice has health issues. By utilizing k-anonymity, Alice would be indistinguishable from at least k – 1 different customer, such that the customer couldn't link the actual request to Alice. As a result, it is necessary that all k customers of the calculated anonymization set sent to the customer have the same obfuscation [4] region such that the customer can't connect the issued position to the home location of Alice.

4) Mix-Zones: Pseudonym is used to break the actual linkage between the user's identities with his/her events. This task is normally performed in most pre-determined areas known as mix zones. The task of the modify is normally performed in most pre-determined areas known as mix zones. A difficulty with this particular method is actually of which there must be adequate customers from mix zone to offer a acceptable level of anonymity.

5) Dummy Locations : This process mostly employs the idea of dummy locations[5] to protect the user's location privacy. A location-dependent issue is actually abstracted as Q = (pos; P), where parameter pos is actually the mobile user location and also parameter P denotes the user specified predicates. We call such a query Q the original query. While using the location dummy strategy, the original problem is typically converted into a query Q0 = (pos1; pos2;:::; posk; P), where the pos1 include the user's real location and k- 1 dummy locations, and P is the original query predicate that applies to all k-locations. We call query Q0 a location privacy query, since it hides the user location.

B. Trajectory Privacy Protection

Some of the existing trajectory privacy protection schemes are as follows -

1) Dummy Trajectory Obfuscation: Protecting trajectory privacy from a data publication viewpoint is performed with simple dummy trajectories obfuscation approach. This approach proposes to generate dummy trajectories so that you can confuse the adversaries. In order to confuse fake trajectories as well as the true ones, dummy trajectories are usually generated under two rules: first, the movement patterns of dummy trajectory needs to be similar to end users; second, the intersections of trajectories needs to be as more as possible. According to these rules, dummy trajectories are usually generated by rotating true users' trajectories. But the main drawback is to generate similar looking trajectories as the quality of anonymity depends upon it.

2) Suppression-Based Method: It is based on the assumption that various adversaries may have diverse and disjoint part of users' trajectories. Suppression-based method decreases the probability of exposing the whole trajectories. Trajectory pieces should be suppressed, publication of these pieces may raise the whole trajectory's breach probability over a particular threshold. This technique works well by preventing the explosion of whole trajectories from the adversaries. But the main setback is that some useful data may get lost during suppression of trajectory data.

3) Trajectory K-Anonymity: Trajectory k-anonymization [6] technique proposes a scheme where every trajectory is generated such that a user finds it indistinguishable to guess the other k-1 trajectories. In this approach first, trajectories are clustered based on log cost metric, then each sample location on trajectories is generalized to a region containing at least k moving objects. Then trajectories are reconstructed by arbitarily choosing sample points from the anonymized region.

4) Trajectory Privacy Preserving Framework: This technique proposes the use of Mix Zone Graph Model where mix zone is applied over a single sensitive location. Pseudonyms of the participators are changed in this mix zone in order to protect the trajectories of the user's which can be inferred by an adversary. Directed Weight Graph of mix zone model is created where an adversary cannot map an exact relationship between participator's arrival time and their exit time. This technique works well but considers only single sensitive location where mix zone graph model is applied.

C. Comparison of Existing Techniques with Proposed System:

Several work exists where location of the users' as well as their trajectories are given privacy. Dummy location [7] is a mechanism of creating fake alias location of the user's

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

location in order to confuse the adversary. Location k anonymity is defined in [8] as a privacy approach designed to protect identification of an individual against a specific datasets. Another technique used for location privacy is obfuscation [9] where the user's location is purposefully altered to lower the precision of the user's spatial temporal information. This can be achieved using generalization or perturbation. Pseudonym [10] is a randomly generated unique identifier provided to each user before entering the sensitive area called as mix zones [11]. Mix zone is the area where a participators movement cannot be tracked by anyone. Pseudonym is generated to break any link present between the user's identity and their events. Mix networks [12] are used to anonymize the channels used between the links between the reports submitted by the user to the system. It is been observed that once a user's trajectory has been identified, then it becomes easy to derive the locations of the users.

Trajectory privacy schemes exist in the literature. Some of them are as follows - dummy trajectories [13] where fake user location trajectories of the users are created. This technique provides privacy to the trajectories however the main problem is how to generate the exact look alike fake trajectories. Another technique proposed is suppression based [14] technique where the whole trajectories are generally suppressed with the assumption that the adversary would not be able to infer the user's information since the whole trajectories are not exposed. The main threat to this approach is that essential data may get lost during the process of suppression. Trajectory k-anonymization [15] technique proposes a scheme where every trajectory is generated such that a user finds it indistinguishable to guess the other k-1 trajectories. All these techniques deal with the whole trajectory and thus increases the storage space cost. Not all locations are sensitive, so providing privacy around these sensitive locations can only be considered instead of whole trajectories [16]. To overcome the defects above, a new scheme has been proposed to preserve the privacy of the trajectories at multiple sensitive locations.

III. EXISTING SYSTEM

Most of the existing techniques focus on providing location privacy of the participators while few approaches consider preserving the trajectories of the user. An approach called as Trajectory Privacy Preserving Framework TrPF for participatory sensing applications has been proposed. The participators known as data collector sense the spatial-temporal information through their mobile device. This information is stored by the Report Server which generates data reports that are eventually stored on Application Server. Any authorized end user or participator can view these reports. Trusted third party severs are used for maintaining security to end users or the data collectors. Fig.1 shows the overall architecture of TrPF system.

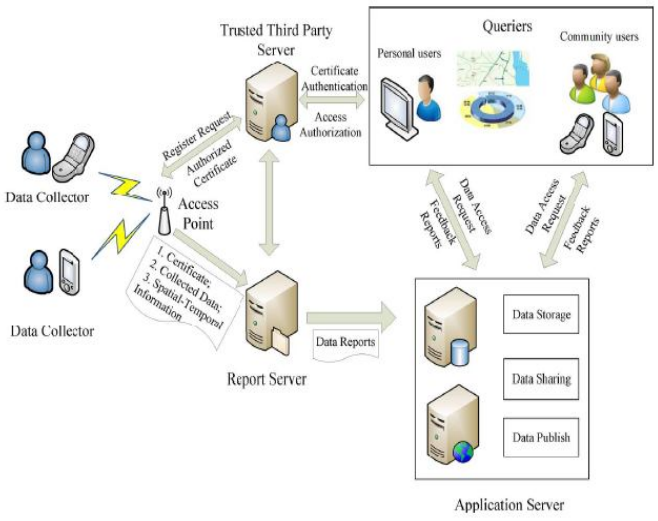


Fig. 1. Architecture Of TrPF System

In this approach the trajectories of the participators are preserved using Mix Zone Graph Model. Not all, but only sensitive trajectories are considered for while applying mix zone graph model. Firstly, a sensitive location, o, is taken as centre and a sensitive area is constructed around it. The trajectories intersecting the sensitive area are said to be as sensitive trajectory segments. Mix zone graph model is then applied on these segments. Thus, the trajectories of the participators are preserved.

A. Limitations

- Only single sensitive location is considered.
- Requires more time to process query as only raw trajectories are considered.

IV. PROPOSED SYSTEM

A. Architecture

The existing solution considers only a single sensitive location while constructing mix zone graph model. This leads to the lack of a systematic approach for global privacy protection. Henceforth to overcome this drawback, the proposed system defines multiple sensitive locations around which multiple mix zone graph model will be applied. Not all point of interests can be considered as the candidates for applying mix zone graph model. The main reason for this the available cost constraints which eventually limits the number of mix zones that one could deploy. So the problem is to address the multiple mix zone graph model's placement. This is an optimization problem.

The proposed system can be explained as follows which is shown in Fig.2 – Firstly, the data collectors sense and provide their spatial temporal information to the Server using their

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

mobile phones. Consider a participator provide their current location(x,y) using GPS embedded in mobile phones. As the participator moves his/her locations get stored on the server, eventually forming location traces i.e trajectories. These trajectories must be preserved from an adversary in order to the preserve the privacy of the participators. Mix Zone Graph Model technique has been used for providing privacy to the participator. In proposed system we consider multiple locations as the candidates for applying Mix Zone Graph Model. Due to cost constraints, not all locations can be considered as point of interests where the model can be applied. Hence selection and placement of multiple locations to be considered for applying Mix Zone Graph Model is the problem to be addressed whose solution is given next. After receiving multiple locations as an output of Multiple Mix Zone Placement Model, Mix Zone Graph Model is applied at all these locations. Meanwhile, an end user may query on this data store on the server and server may provide appropriate result.

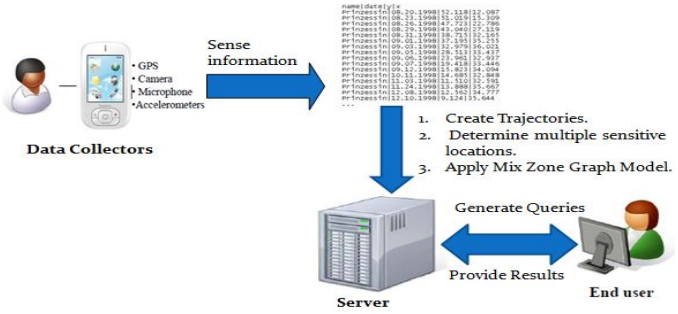


Fig 2. Proposed System Architecture

For instance, consider a scenario of Online Car Booking System where end users i.e the customers at any time can book a car online through the system. The administrator depending on the availability of the drivers assigns a driver to the customer. The customer at any time can track their assigned driver. Considering the privacy of the driver, not all location trajectories of the driver should be visible to the customer. The driver who is the participator in this system provide their locations to the administrator using their mobile phones. Trajectories of driver's are stored on the server which can be viewed by the administrator and driver itself. No the other party should be able to view the whole trajectory of a driver. Hence, Mix Zone Graph Model is applied at multiple sensitive locations of the driver thereby not allowing the customer to view the whole trajectory of a driver. The sensitive locations of the driver like his house, hospital, gym, work place, etc. must not be able to be known by the customer or an adversary. Applying Mix Zone Graph Model at multiple locations prevents an adversary or an end user from inferring the whole trajectory of the participator. Thus the trajectory privacy of the participatory sensing user is preserved.

B. Solution - Multiple Mix Zone Placement Model

This approach generally determines the number of positions where Mix Zone Graph model has to be applied. Basically this approach first finds the points (vertices), whose removal makes the graph disconnected. Such points are called as articulation points. This partitions the graph into disconnected components thus eliminating the need of pair wise connections between them. To refine the quality of solution further, the set of independent vertices are found. These are the vertices that are not adjacent to each other. Finally, the number of mix zones are limited by the given cost constraint.

Consider the following Graph G = (V,E) where vertices V represents Points Of Interests of a participator and E represents the road segments connecting POIs. The first step is built on the observation that partitioning G into several disconnected components is helpful to eliminate the pairwise connections across these components. Therefore, we are seeking for vertices whose removal disconnect the graph. Such vertices are typically referred to as articulation points in graph theory. Take the area graph in Figure 3 as an example. Any route from 1 to 9 or from 1 to 12 needs to go through vertices 6 and 10. Therefore, 6 and 10 are articulation points in this graph. If a mix zone is deployed at vertex 6 or 10, a pseudonym appears at any vertex in the bottom part of the graph cannot appear at vertices 9, 12, and 11. Hence, the total number of pairwise associations is reduced.

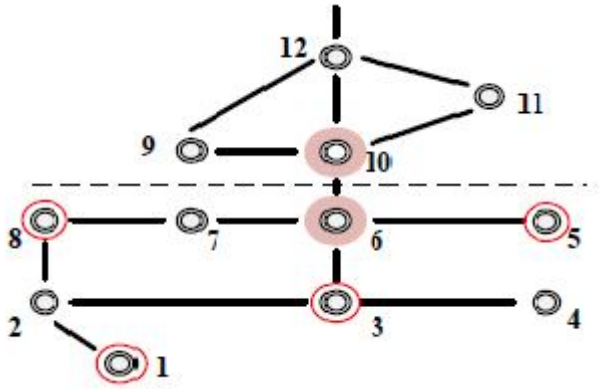


Fig.3 Point of Interests Graph

After G is partitioned into disconnected components, the mix zone deployment in each component is further refined to improve the solution quality. In graph theory, an independent set refers to a set of vertices that are not adjacent to each other. Hence, if all vertices that are not in an independent set are selected as mix zones, there will be no pair wise association between the vertices in the independent set. Again, refer to the bottom part of Figure 3 as an example. Circle highlighted vertices, {1, 8, 3, 5}, form a maximal independent set for the lower part of the graph. If vertices {2, 4, 6, 7} are selected as mix zones, a user Alice's pseudonym ux appears at vertex 1 will not appear at any other vertex in the independent set. As a result, Alice's past and future locations on her trajectory are protected, even though her identity gets exposed at vertex 1. Finally, there is a need to control the number of mix zones to meet the cost and service constraint. At the last step of our algorithm, we iteratively remove the vertex that introduces the

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

least number of pair wise association increment from the mix zone candidate set selected by previous steps until cost constraint is met.

C. Mathematical Model

Let **S = {I, P ,O}**
I = Input
O = Output
P = Process.

I= {SI, GQ}
SI = Sense Information
GQ= Generate Query

P = {TR, MMPM, MZGM}
TR = Generate trajectories.
MMPM = Determine multiple locations.
MZGM = Apply Mix Zone Graph Model.

O = {PR}
PR = Provide results of the generated queries.

Fig. 4 represents the mathematical model of overall proposed system.

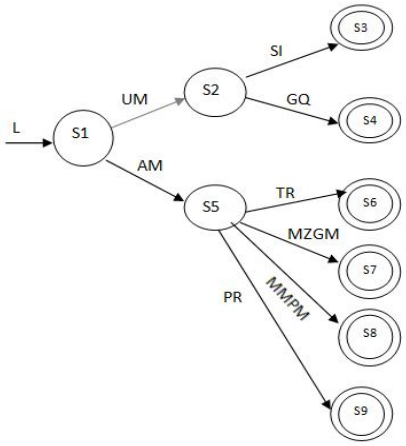


Fig.4. Mathematical Model

L: Login into the system.
UM: Access to the User Module.
AM: Access to the Administrator Module.
SI : Sense the Information
GQ: Generate Queries.
TR: Generate trajectories.
MZGM: Generate Trajectory Mix Zone Graph Model.
MMPM: Generate Multiple Mix Zone Placement Model.
PR : Provide Results to the user.

D. Algorithms Used

1) GraphConstruct Algorithm

This algorithm is used to construct mix zone graph model which is represented by a graph G(V,E). A mix zone graph model has been proposed such that Directed Weight Graph (DWG) is represented by G = (V, E) , where,
V represents set of vertices that are constructed as the pseudonyms.
E represents set of edges that represent the participators' trajectory mapping from the ingress to the egress in the sensitive area.

Algorithm1 GraphConstruct

Input :- Trajectory Tr and pseudonym set P.

Output :-Directed Weight Graph (G).

1 : Procedure
2: Define sensitive location and construct sensitive area around it such that Si = {o, r} where o is sensitive location and r is the radius.
3: Determine the set of sensitive trajectory segments Tf.
4: Randomly select ingress pseudonym Pi and assign it to the vertex Vi.
5: Randomly select egress pseudonym Pj such that Pj $\neq$ Pi and assign it to the vertex Vj.
6: Construct Edge Eij such that Eij -> (Vi,Vj)
7: Assign weight Wij to each edges using Weight Construct algorithm.

2) Algorithm 2:- Weight Construct

This algorithm is used to find weights of the edges formed in the graph of the mix zone graph model. Here,
Vi represents participator entering the mix zone.
K represents total number of participators entering mix zone.
Pi represents ingress pseudonym of a participator.
Pj represents egress pseudonym of a participator.
tingress(Vi) represents time at which participator enters the mix zone.
tj to tj+1 represents time interval during which participator exists from the system.
P(Vi,t) represents the probability that a single participator exits the mix zones between time interval[tj,tj+1].
The participator Vi generally takes tj –tingress(Vi) to tj+1 – tingress(Vi) time in mix-zone for data collection.
Δ' **t** represents data collection time in mix zone.

is the probability density function (PDF) of data collection time in mix-zones.

Therefore,

P(Vi, t) = $\int_{tj-tingress(V}^{tj+1-tingres}$ (1)

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

The above mentioned equation represents probability of a single participator exiting from the mix zone. Thus, the probability for all the participators exiting from the mix zone is given by (2)
P(V', t) represents the probability that all participator exits from the mix zone between time interval[tj,tj+1].

P(V' , t) = (2)

However only one of them is a real participator. Hence, the probability that the participator Vi exits in the time interval [tj,tj+1] is denoted by P(Vi [tj ,tj+1]) is given by the following conditional probability-

P(Vi [tj ,tj+1]) = (3)

Wij is given by Wij= P(Vi [tj ,tj+1]) such that wij is between 0 to 1 and i [1,k] and

(4)

The Weight Construct algorithm is given as follows:-

Algorithm 2 WeightConstruct

Input :- tingress and Δtegress=[tj,tj+1] and Δ' t

Output:- Edge Weight W

1: Procedure
2: Determine the probability P(Vi,t) of single participator exiting mix zone in given time interval.
3: Determine the probability P(Vi',t) for all participator exiting mix zone in given time interval.
4: Find the probability of a single participator exiting the mix zone model denoted by **P(Vi [tj ,tj+1])**
5: Assign **P(Vi [tj ,tj+1])** as the weight of the edge.

3) Algorithm 3:- Multiple Mix Zone Placement Model.

This algorithm generally determines the number of positions where mix zone graph model has to be applied. Basically this algorithm first finds the points (vertices), whose removal makes the graph disconnected. Such points are called as articulation points. This partitions the graph into disconnected components thus eliminating the need of pair wise connections between them. To refine the quality of solution further, the set of independent vertices are found. These are the vertices that are not adjacent to each other. Finally, the number of mix zones are limited by the given cost constraint.

Algorithm 3 Multiple Mix Zone Placement Model.

Input :- A graph G and Z.

Output:- A set of at most NP selected mix zone positions

1: Procedure
2: Find articulation points in the given graph.
3: Find maximal independent set.
4: Maintain Cost Constraint.

V . EVALUATION AND RESULTS

Here, the real time data is taken as an input for the system. As explained prior, the participator is providing their location (x , y) and timestamp (t) to the Server using the GPS of their mobile phones. The trajectories are stored in the form of (tid, x, y ,t) on the server where tid represents the trajectory ID. Sensitive locations are considered around which Mix Zone Graph Model is applied. Meanwhile the end user can access the relevant data on the server . The Online Vehicle Booking System is built as an website using C# ASP.Net whereas the participator provides its spatial temporal data to the Server using Android mobile devices. Participator side module is developed using Android Programming in Java.

This work aims in proving that the privacy level of a participator can be enhanced by applying Mix Zone Graph Model at multiple sensitive locations instead on single sensitive location. This can be proved by measuring the rate of successful attacks on single mix zone as compared to multiple mix zones. An attack is successful if the adversary finds out the corresponding pseudonym used by a user in the side information. The success rate of an adversary is the ratio of number of successful attacks over total number of attacks. Fig.5 shows the attack success rate when different number of mix zones is applied where X axis represents number of mix zones to be deployed at various sensitive locations and Y axis represents the rate of successful attack.

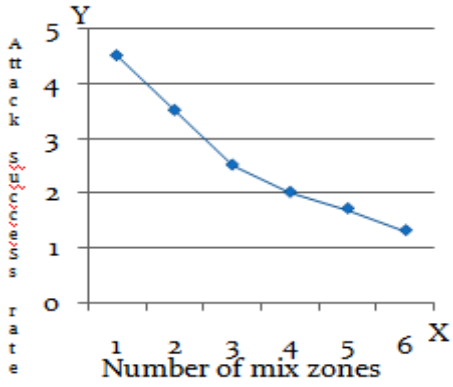


Fig.5. Comparison of Privacy Level

The above graph shows that the rate of successful attack is high when number of mix zones is less. It shows that as the number of mix zones increases eventually the rate of successful attack decreases thereby improving the level of privacy. The reason for this is on increase in number of mix

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

zones successful attack rate decreases because the adversary has to crack the corresponding number of pseudonyms in order to deduce the whole trajectory. This becomes sustainably simpler for an adversary with single mix zone as only one pseudonym has to be cracked. So as the number of locations where mix zone graph model has to be applied increases, the privacy preservation of trajectories increases. Thus, the proposed scheme offers better privacy as compared to the existing systems.

Another advantage of the proposed work is that it requires less storage space as compared to the existing techniques. Previous work like Dummy trajectories and trajectory k-anonymity stored all trajectories for providing protection. Given t trajectories and each trajectory contains N segments then the storage space required will be O(N* t) to store total t trajectories. Whereas trajectory mix zone graph model approach requires only pseudonym to be stored. Only sensitive trajectory segments are considered here and not all trajectories. Hence storage space required for this approach is quite less as compared to the previous work. Further, the increase of trajectories may not affect the number of pseudonyms too much. By comparison, our proposal has lesser storage memory than that of the other proposals.

VI. CONCLUSION AND FUTURE WORK

Participatory sensing leverages the ubiquity of mobile phones to open new perspectives in terms of sensing. The analysis has revealed that virtually all applications capture location and time information. The collected data is been stored in form of the trajectories. The privacy of these trajectories needs to be preserved. Trajectory Mix zone Graph model is been used here for providing privacy to the trajectories of the participators'. This approach proposes multiple sensitive locations to be considered for applying Mix Zone Graph Model as opposed to single sensitive location. The results proves that applying mix zone graph model at multiple sensitive locations as compared to single sensitive location increases the privacy level of the participator. Hence the proposed system provides better results as compared to the existing techniques in terms of increased privacy level and reduced storage space. In future, mix zone graph model can be applied on multiple sensitive locations of semantic trajectories.

ACKNOWLEDGMENT

The authors appreciate the helpful comments and suggestions from the anonymous reviews.

REFERENCES

[1] T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A.Peterson, "People-centric urban sensing," *in Proc. 2nd Ann. Int. Workshopon Wireless Internet*, 2006, p. 18, ACM.

[2] Sheng Gao, Jianfeng Ma, Weisong Shi, Senior Member, IEEE, Guoxing Zhan, and Cong Sun, "TrPF: A Trajectory Privacy-Preserving Framework for Participatory Sensing" *IEEE transactions on Information Forensics and security*, vol. 8, no. 6, June 2013.

[3] X. Liu, H. Zhao,M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," *in Proc. IEEE INFOCOM*, 2012, pp. 972-980.

[4] M. Duckham and L. Kulik, "A formal model of obfuscation and negotiation for location privacy," *in Proc. 3rd Int. Conf. Pervasive Computing (PERVASIVE'05)*, 2005, pp. 152–170.

[5] H. Lu, C. S. Jensen, andM. L. Yiu, "Pad: Privacy-area aware, dummy based location privacy in mobile services," in Proc. 7th *ACM Int.Workshop on Data Engineering for Wireless and Mobile Access*, 2008, pp. 16–23, ACM.

[6] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," in *Proc. ACM SIGSPATIAL ACM GIS 2008 Int. Workshop on Security and Privacy in GIS and LBS*, 2008, pp. 52–61.

[7] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in Proc.Int. Conf. Pervasive Services, 2005, pp. 88-97.

[8] L. Sweeney, "k-anonymity: Amodel for protecting privacy," Int. J. Uncertainty Fuzziness and Knowl. Based Syst., vol. 10, no. 5, pp. 557-570,2002.

[9] M. Duckham and L. Kulik, "A formalmodel of obfuscation and negotiation for location privacy," in Proc. 3rd Int. Conf. Pervasive Computing(PERVASIVE'05), 2005, pp. 152-170.

[10] J. Freudiger, M. H. Manshaei, J. Y. Le Boudec, and J. P. Hubaux, "On the age of pseudonyms in mobile ad hoc networks," in Proc. IEEEINFOCOM, 2010, pp. 1-9.

[11] J. Freudiger, M. Raya, M. Flegyhzi, P. Papadimitratos, and J. P. Hubaux, "Mix-zones for location privacy in vehicular networks," in Proc. 1st Int. Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS 07), Vancouver, BC, Canada, 2007.

[12] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, and D. Kotz, "Anonysense: Opportunistic and privacypreserving contextcollection," Pervasive Comput., vol. 5013, pp. 280-297, 2008.

[13] T. H. You, W. C. Peng, and W. C. Lee, "Protecting moving trajectories with dummies," in Proc. IEEE Int. Conf. Mobile Data Management,2007, pp. 278-282.

[14] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in Proc. IEEE 9th Int. Conf. Mobile Data Management (MDM'08), 2008, pp. 65-72.

[15] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," in Proc. ACM SIGSPATIAL ACM GIS 2008 Int. Workshop on Security and Privacy in GIS and LBS, 2008, pp. 52-61.

[16] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in Proc. 2008 ACM Symp. Applied Computing 2008, pp. 863-868, ACM.

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

Security Analytics in Big Data Infrastructures

K.BOUKRI* and H.CHAOUI**

Systems Engineering Laboratory, the ADSI team
National School of Applied Sciences in Kenitra. Campus Universitaire, BP 241 Kenitra Morocco

ABSTRACT

Big Data is usually so large and complex, and it has become an emerging hot topic in network security fields. How to deal with lots of safety data which are produced by heterogeneous network security devices, how to analyze and coordinate such Big Data network events must be studied to date [1-3]. This paper develops a kind of Big Data Analytics Security framework using collection, analysis and integration, event-correlation and scenario analysis technique to process the raw data gathered from Big Data infrastructure (BDI).

Keywords: Big Data Security, Big Data analytics, Intelligence Analysis, Intrusion detection

I. INTRODUCTION

As Big Data systems play increasingly vital roles in modern society, the security of Big Data behavior become more and more important. The relevant security techniques such as firewall, anti-virus, VPN, IDS and Security audit have been developed to protect against the security threats. However, much more issues are exposed in the practice of security implementation and deployment. There is a lack of an effective method to analyze the security events generated by various network equipment in BDIs.

The use of multiple and diverse sources producing huge amounts of data calls for the research of new solutions for monitoring and analysis, able to timely and efficiently recognize ongoing malicious activities in BDIs. The age of Big Data, network security monitoring systems need to meet several requirements. First, we must find abnormal alarm as soon as possible. This can take early measures to avoid or mitigate the impact on services. This is actually an application of trend forecasting. Secondly, we must make a correct diagnosis of alert information from the massive, extract the real, non-redundant information, in order to find out the root of the problem, and to solve the problem.

Sensitivity and reliability is a pair of contradictory of network security monitor in Big Data environment, and at the same time, the detection system's network level and performance decide their information acquisition and analysis limitations. We obtained information on these alerts, it is inevitable there will be some missing, and it can be applied to Big Data missing information predict, to get a true picture of network attacks or abnormal.

Big Data can give full play to its externality and generate much larger than the sum of the huge value through fusion with some related data cross. For network security monitoring in Big Data environment, the most important step is the association of events. There are many researches on traditional event correlation, such as rule-based correlation, Bayesian network inference, model-based reasoning, and filtering, case-based reasoning artificial neural network reasoning.

In this context we propose a framework comprising visual analytics components which are coordinated by modules working on behalf of a security analysis. This paper is structured as follows: In Section II, we present the big data analytics security needs and challenges. In Section III, we explain our proposal framework design to Big Data Analytics for Security (BDAS). We conclude the paper and present future work in Section IV.

II. NEEDS AND CHALLENGES

At present, there are more and more attention and research on network security monitoring, such as intrusion-detection based on data analysis and integrated management of network security [4, 5]. However, most current works are trying to research and develop a network security monitoring system based on partial view of the whole network. But we still found following outstanding issues during the practice of threats detection in these systems:

1. During the inchoate construction of BDIs, in various security equipment, there is a lack of communication ability of management and security

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

information with others. This shortage reduces the holistic system efficiency and increases the cost of discovery-time and response-time.

2. The detection and reporting of security events only present some kind of raw data but without more "information" and "knowledge", as the result of lacking enough analysis ability.

3. Though there are lots of researches on event-correlation, such as rule-based correlation, Bayesian network reasoning, model-based reasoning, filtering, case-based reasoning and artificial neural network reasoning, etc. But in most cases, these techniques are only applied to single abnormity detection system, and also not focused on holistic BDI, including the network equipment, and sub-domain hosts, etc. Furthermore the deficiency of rapid and effective communication mechanism dissevers the relationship among monitoring components. Based on above analysis, this paper develops a framework for big data analytics security, which can highly detect anomalies in BDIs.

III. PROPOSAL FRAMEWORK FOR DATA ANALYTICS SECURITY

A. The goal of the proposal framework

More recently, the focus of security has shifted to monitoring network and Internet traffic for the detection of bad actions as compared to the traditional approach of the detection of bad signatures. Specifically, traditional security is focused on catching malware by scanning incoming traffic against malware signatures which only detect limited-scope threats that have been already encountered in the past. In addition, the development of signatures lags far behind the development of attack techniques. Thus, techniques like intrusion detection systems, firewalls and anti-virus software can be easily rendered ineffective by attackers. This scenario has become more crucial in the presence of big data within computer networks – petabytes and Exabyte of information being transferred daily between nodes make it very easy for attackers to enter any network, hide their presence effectively and cause severe damage efficiently. These big data problems are stressed in the following points:

- Corporations are now extending their data networks to allow partners and customers to access data in different ways to facilitate collaboration, hence making networks more vulnerable to attacks. The advent and extensive use of cloud and mobile computing have also generated new attack methods.
- The advent of big data has seen a corresponding increase in the hacking skills of attackers, and evading traditional security measures such as signature-based tools, which is now a thing of the past.
- Due to big data, it is possible to collect a relatively small slice of security information, e.g. network logs, Security Information and Event Management (SIEM) alerts, access records etc. Hence, damage done by new intrusion methods could be realized only after an attack.
- Big data also prevents most security data from being analyzed due to its complexity, e.g. data could be coming from different sources. It could be stored in different formats on different machines, or could be generating too quickly to make any type of analysis feasible through traditional techniques, computer hardware and software architectures.

Security Analytics address these issues by reinventing the wheel of Big Data security. It employs techniques from Big Data Analytics (BDA) to derive useful information for preventing attacks [6]. It provides the following unique features:

- A more agile decision-making approach for networks managers with surveillance and monitoring of real time network streams,
- Dynamic detection of both known and previously unknown suspicious or malicious behavior, usage access pattern, transaction and network traffic flow, applicable to all types of intrusion threats
- Effective detection of suspicious and malicious behavior (least possible false positive rate),
- Ability to deal with suspicious and malicious behavior in real-time,
- Appropriate dashboard-based visualization techniques to provide full visibility (360° view) of network progress and problems in real-time.
- Appropriate big data hardware and software to cope up with the aforementioned requirements.

B. Framework Model

We have developed a framework for Big Data Analytics Security based on above considerations.

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

1. DATA PRE-PROCESSING

This module collects the raw data from the general data source in Big Data system, including the flow-rate, event logs and security logs from the data source. Most data collected are disorder in hybrid formats and descriptions, which makes the system behavior detection work much difficult. We use the Data Standardizing and Integrating Module to solve this problem. Firstly, this module standardizes the raw data into unified format, using the pre-defined pattern. Next, it filters the standardized data and removes redundancy from it. Then, it classifies the data by following the classification rules.

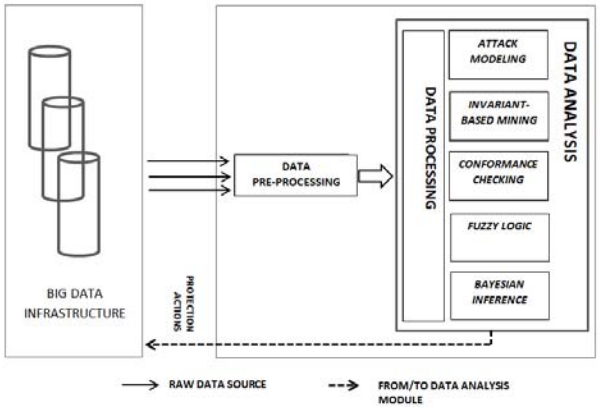


Figure 1 : The Framework for Big Data Analytics Security

2. DATA ANALYSIS

This component analyzes the data and provides as outputs information on (i) how to adapt the grain monitoring, (ii) the actions of protection should be carried out on the architecture. Starting from our past experiences on attack modeling and data analysis, we consider the following functional blocks.

Data Processing: it allows the collection of raw data usually contain unnecessary or redundant information which may affect the effectiveness of the analysis [7]. The first step in the analysis is to be performed to retrieve raw data, the adoption of the filtration technique or coalescence event, such as those analyzed in [8].

Attack Modeling: this function block provides tools to define and statically analyze patterns of attack. The driver model used in this block must be able to: (i) providing a high degree of flexibility in the representation of many different security scenarios compactly; (ii) enable the specification of different types of constraints on the possible timing attacks; (iii) representing attack scenarios at different levels of abstraction, to focus the task of compliance in various ways. Typed temporal graph-based attack models [9] seem to be good options for the above requirements. They are rich in terms of time constraints that can be expressed. In addition, it is relatively easy to manipulate the definition of hierarchies of generalization / specialization between different types of events.

Conformance Checking: the main purpose of this functional block is the detection of cases of attack, in sequences of events recorded by the compliance behavior connected with the given set of attack patterns. The main requirement of this block is scalability. In real-world scenarios, the protection of critical infrastructure, are available on the online system, where we would like to raise an alert as soon as when an event with a "criticality" above the threshold is connected. It is therefore important to define appropriate data structures, to ensure rapid access to relevant information and appropriate algorithms that are closely associated with such structures to ensure rapid detection of an attack [10], [11]. Furthermore, it is important to identify the conditions that make the problem tractable from a theoretical point of view.

In fact, recent work on case detection automata as temporal models in the recorded sequence of events [12], [13], [14] has shown that the detection time if appropriate in the real world can be obtained by limiting the number of partial solutions through a form of temporal filtering based on constraints. Finally, the parallelization two data structures and algorithms for checking compliance (see [15]) seem mandatory when we are targeting large data for the protection of security.

Invariant-based Mining: invariants are properties of a system that are responsible for keeping all its executions. If those properties are to be violated (or broken) while monitoring the execution of the system, it is possible to trigger alarms useful to undertake immediate protective actions. Several studies have confirmed that it is possible to discover invariants of complex real-world systems [16], [17]. However, in our case, the challenge is to find invariant relationships in large data collected from the architecture. The exploration of base-invariant block intends to deal with this problem, while performing two tasks: i) automatic extraction of invariants flow data using autoregressive models and ii) detection at runtime when the invariant relationships are broken, to trigger

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

immediate action. A preliminary application of the approach on real data collected from a cloud software system production has demonstrated its feasibility and usefulness to discover the deviation of performance and SLA violations. [18]

Fuzzy Logic: statistical methods cause a lot of false alarms. This is due to the difficulty of defining precise and crisp rules describing when an event is an anomaly or not. The boundaries between the normal and the abnormal behavior of a system are not clear, and deciding the degree of intrusion at which the alarm to be raised may vary in different situations [19]. Fuzzy logic is derived from fuzzy set theory to process the approximate rather than precise data reasoning and contributing effectively to facilitate the abrupt separation of normality and abnormality. [20] The degree of truth of an expression is not clear, and the use of fuzzy linguistic variables can express imprecision in measurement. In certain embodiments, a percentage of 99.95% for the attack detection accuracy was achieved [19].

Bayesian Inference: security monitors usually produce a large number of false alarms. A Bayesian network approach can be used on the top of the architecture for correlating alerts from different sources and to filter false notification. This approach has been successfully used to detect attacks flight credentials. [21] Raw alerts generated during the progression of an attack, such as violations of user IDS profile and notifications are correlated through a Bayesian network to identify users by hijacking compromise. The approach was able to eliminate about 80% false positives (A user is not compromised declared compromise) without missing any compromise user.

IV. CONCLUSION & PERSPECTIVES

Traditional security solutions are not capable anymore of encompassing the real-time big data network streams using traditions tools and techniques. We have shown how Security Analytics (the application of Big Data Analytics techniques to derive actionable intelligence and insights from streams in real-time) is rapidly becoming a strong need for Big Data security setups. Although the current adoption of analytical solutions is by no means revolutionary, the awareness of adoption is increasing rapidly. To support this cause, in this paper, we specifically mention the needs and challenges in security analytics. We next highlight the goal of our proposal framework to address data-driven security analytics issues. We then describe the main components in our proposal framework, which is design to apply security analytics techniques BDIs in order to decrease the false positive rate of a predictive model for many attacks. In our future works, we plan to implement our framework in a real infrastructure in other to evaluate and test the efficiency of our proposal.

REFERENCES

[1] James Manyika, Michael Chui, Brad Brown and all: Big data: The next frontier for innovation, competition, and productivity. 2011.

[2] Nature. Big Data[EB/OL]. [2012-10-02]. http://www.nature.com/news/specials/bigdata/index.html

[3] Chen Y, Alspaugh S, Katz R. " Interactive query processing in big data systems: A cross-industry study of Map Reduce workloads". [J].PVLDB, 2012, 5(12).pp.1802-1813.

[4] Kemmerer R.A., Vigna G, "Intrusion detection: a brief history and overview", IEEE Computer, Vol 35(4),pp. 27-30, 2002.

[5] Thottan. M., Chuanyi Ji. "Anomaly detection in IP networks". Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on], 2003, 51(8) 2191 2204.

[6] A. Sathi, "Big Data Analytics: Disruptive Technologies for Changing the Game", Mc Press, 1st Edition, February 5, 2013.

[7] A. Pecchia and S. Russo, "Detection of software failures through event logs: An experimental study", in Software Reliability Engineering (ISSRE), 2012 IEEE 23rd Intnl. Symp. on, Nov 2012, pp. 31–40.

[8] C. Di Martino, M. Cinque, and D. Cotroneo, "Assessing time coalescence techniques for the analysis of supercomputer logs", in Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on, June 2012, pp. 1–12.

[9] A. Guzzo, A. Pugliese, A. Rullo, and D. Sacca, "Intrusion detection with hypergraph-based attack models", in GKR, ser. Lecture Notes in Computer Science, M. Croitoru, S. Rudolph, S. Woltran, and C. Gonzales,Eds., vol. 8323. Springer, 2013, pp. 58–73.

[10] M. Albanese, A. Pugliese, V. S. Subrahmanian, and O. Udrea, "Magic: A multi-activity graph index for activity detection", in IRI. IEEE Systems,Man, and Cybernetics Society, 2007, pp. 267–272.

[11] M. Albanese, A. Pugliese, and V. S. Subrahmanian, "Fast activity detection: Indexing for temporal stochastic automaton-based activity models", IEEE Trans. Knowl. Data Eng., vol. 25, no. 2, pp. 360 373, 2013.

[12] M. Albanese, S. Jajodia, A. Pugliese, and V. S. Subrahmanian, "Scalable analysis of attack scenarios", in ESORICS, ser. Lecture Notes in Computer Science, V. Atluri and C. D´ ıaz, Eds., vol. 6879. Springer,2011, pp. 416–433.

[13] N. Chaki and A. Cortesi, “Scalable detection of cyber-attacks”, in CISIM, ser. Communications in Computer and Information Science, ,Eds., vol. 245. Springer, 2011, pp. 9–18.
[14] M. Albanese, A. Pugliese, and V. S. Subrahmanian, “Fast activity detection: Indexing for temporal stochastic automaton-based activity models”, IEEE Trans. Knowl. Data Eng., vol. 25, no. 2, pp. 360 373, 2013.
[15] A. Pugliese, V. S. Subrahmanian, C. Thomas, and C. Molinaro, “Pass:A parallel activity search system”, IEEE Trans. Knowl. Data Eng., to appear.
[16] A.B. Sharma et al., “Fault detection and localization in distributed systems using invariant relationships”, in Proc. of 43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN),2013.
[17] J.-G. Lou et al., “Mining invariants from console logs for system problem detection”, in Proc. Of the USENIX annual technical conference, ser. USENIXATC’10, 2010.
[18] S. Sarkar, R. Ganesan, M. Cinque, F. Frattini, S. Russo, and A. Savignano, “Mining invariants from saas application logs”, in Tenth European Dependable Computing Conference (EDCC 2014), May 2014.
[19] A. Feizollah, S. Shamshirband, N. Anuar, R. Salleh, and M. Mat Kiah, “Anomaly detection using cooperative fuzzy logic controller”, in Intelligent Robotics Systems: Inspiring the NEXT, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2013, vol. 376, pp. 220–231.
[20] F. Frattini, M. Esposito, and G. Pietro, “Mobifuzzy: A fuzzy library to build mobile dsss for remote patient monitoring”, in Autonomous and Intelligent Systems, ser. Lecture Notes in Computer Science, Eds. Springer Berlin Heidelberg, 2012, pp.79–86.
[21] A. Pecchia, A. Sharma, Z. Kalbarczyk, D. Cotroneo, and R. K. Iyer,“Identifying compromised users in shared computing infrastructures: A data-driven bayesian network approach”, in Proceedings of the International Symposium on Reliable Distributed Systems (SRDS). IEEE Computer Society, 2011, pp. 127–136.

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

Switching of code between the users for Enhancing the Security in OCDMA

Sumit Gupta, Aditya Goel

Abstract: A new technique is proposed for optical code division multiple accesses for enhancing the security against the eavesdropper. Switching of the pulse spectrum of code is performed between the two users and switching position of the pulse spectrum of code varies from group to group. In this technique every pulse of a code does not have direct information. Code detection probability of individual user decreases against the eavesdropper. The analysis and simulation result compares with the exiting method MQC, RD and MDW code.

Key Words: Random Diagonal (RD) Code , Modified Double Weight (MDW) Code, Zero cross correlation code (ZCC),Spectral Amplitude Coding (SAC)

I. INTRODUCTION

In optical networks, the Optical code division multiple access (OCDMA) is getting more and more attraction as multiple user shares the communication network asynchronously and synchronously with high security level [1-2].In optical code division multiple access system, user information is transmitted by assigning the code address on OOK (on off keying) based pattern to '1' information bit. Information about a user is extracted at the receiver end by correlating the assign code between the users [3-4].Multiple Access interface (MAI) exists in the network due to in phase cross correlation property that degraded the system performance. MAI is reduced by reducing the in phase cross correlation property and detection technique. Zero in phase cross correlation in code of different users, eliminates the effect of the MAI. Security of the network against the eavesdropper is another issue that can be enhanced in OCDMA by increasing the length of code such as Modified Frequency Hopping code, Optical Orthogonal Codes, Modified Double Weight code (MDW) and MQC [8-10]. These codes suffer too long length of code situation and large weight problem. This requires the broadband source and narrow band filters.

Sumit Gupta is research Scholar in Electronics and Communication Engineering Department Maulana Azad National Institute of Technology Bhopal,India
Aditya Goel is Professor in Electronics and Communication Engineering Department Maulana Azad National Institute of Technology Bhopal, India

If the highly sensitive detector is used by the eavesdropper than information can be detected by the any spectral pulse of code of weight W, So a new technique is proposed in this paper to remove the MAI and the problem of long length code by switching a part of code between the user. then the information is available with few number of chip in direct manner. This is discussed in section II. Mathematical analysis of the proposed new method is explained in section III. Section IV shows the comparison of different method with proposed method. Paper ends with the conclusion .

II. PROPOSED METHOD

A coding pattern is proposed in table 1.1 for W weight and N number of users with in phase zero cross correlation property. A 2x2switch of defined sequence is used between the spectral pulse Position of code of two users. Pulse spectrum (chip) position for switching between the users varies from group of two users. Switch S_{11}, S_{12} are the inputs of the switch S_1 and Switch S_{21} and S_{22} are the inputs of the switch of S_2.

Probability of code detection of each user if all code chips (Weight) are to be detected than in case of zero in phase cross correlation code eq .1 and in case of unity cross correlation eq.2[8,11]

$$P(U) = \frac{1}{L} \times \frac{1}{L-1} \times \ldots \times \frac{1}{L-W_z} \quad (1)$$

$$P(Z) = \frac{1}{L} \times \frac{1}{L-1} \times \ldots \times \frac{1}{L-W} \quad (2)$$

$$P(S) = \frac{1}{L} \times \frac{1}{\left((L-1)_{C_2}\right)2^N} \quad (3)$$

Where W_z are the weight with zero cross correlation in MDW code

If the high performance detector is used. Each chip of Code carrying the information so Probability of code detection for zero cross correlation code in equation 4 and Probability of code detection in the case of the proposed design in equation 5

$L = WN$ L=Length of code N=number of user

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

$$P(Z) = \frac{1}{N} \quad (4)$$

$$P(S) = \frac{W - W_S}{L} \quad (5)$$

TABLE 1.1 CODE PATTERN WITH SWITCHING OF SPECTRAL PULSE.

Users	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}	λ_{11}	λ_{12}
1	1		1									
2		1		1								
3					1		1					
4						1		1				
5									1		1	
6										1		1
Switch Inputs		S_{11}	S_{12}				S_{21}	S_{22}	S_{31}			S_{32}

III. MATHEMATICAL ANALYSIS OF BER

For analysis of this system we use the Gaussian approximation in our calculation [4,10] . This system is based on the zero cross co –relation so, then we only consider the thermal noise (Rth) and shot noise (Rsn) in respect to PIIN.

Let CK(i) denotes the ith element of K user in this code ZCC than the following assumptions are made.

a. Each light source spectrum is flat over the bandwidth [Vo-ΔV/2 , Vo– ΔV/2] where Vo is central frequency and ΔV is the optical source bandwidth in Hertz.
b. Each power spectral component has an identical spectral width.
c. Each user has nearly equal power at the transmitter.
d. Each user bit stream is synchronized

Each light source spectrum is flat over the bandwidth [Vo-ΔV/2 , Vo– ΔV/2] where Vo

The power spectral density (PSD) of the received signals can be given as

$$r(v) = \frac{P_{sr}}{\Delta v} \sum_{k-1}^{k} d_k \sum_{i-1}^{N} c_k\,(i) rect(i) \quad (6)$$

$$rect(i) = u\left[v - v_0 - \frac{\Delta v}{2L}(-L + 2i - 2)\right] - u\left[v - v_0 - \frac{\Delta v}{2L}(-N + 2i)\right] = u\left[\frac{\Delta v}{L}\right] \quad (6)$$

Where $u(v)$ is the unit step function expressed as:

$$u(v) = \begin{cases} 1, & v \geq 0 \\ 0, & v < 0 \end{cases}$$

$$\int_0^{\infty} G(v)dv = \int_0^{\infty} \left[\frac{P_{sr}}{\Delta v} \sum_{k-1}^{k} d_k \sum_{i-1}^{k} C_k(i) C_I(i) rect(i)\right] dv \quad (7)$$

$$\int_0^{\infty} G(v)\,dv = \frac{P_{sr}}{\Delta v}\left[\sum_{k-1}^{k} d_k\,.W.\frac{\Delta v}{L} + \sum_{k\neq 1}^{k} d_k\,.0.\frac{\Delta v}{N}\right] \quad (8)$$

The value of $\sum_{k-1}^{k} d_k$ is equal to 1 then

$$\int_0^{\infty} G_{dd}\,(v)dv = \frac{P_{sr}W}{L} \quad (9)$$

The photo current I can be expressed as

$$I = I_{dd} = \Re \int_0^{\infty} G_{dd}(v)\,dv \quad (10)$$

The variation of photocurrent due to detection of an ideally un polarized thermal light can be expressed as

$$I = \Re\left[\frac{P_{sr}W}{L}\right]$$

$$\langle I^2 \rangle = 2eB(I_{dd}) + \frac{4K_bT_nB}{R_L}$$

$$\langle I^2 \rangle = 2eB\Re\left[\int_0^{\infty} G_{dd}(v)dv\right] + \frac{4K_bT_nB}{R_L} \quad (11)$$

When all users transmitting 1 than probability of each user sending 1 is ½ than equation .11 become

$$\langle I^2 \rangle = \frac{P_{sr}eB\Re}{L}[W] + \frac{4K_bT_nB}{R_L}$$

The signal to noise ratio of direct detection technique is given by following equation

$$SNR = \frac{\langle I_{dd} \rangle^2}{\langle I^2 \rangle}$$

When putting all equation than new formula for SNR will be

$$SNR = \frac{\frac{\Re^2 {P_{sr}}^2 (W)^2}{L^2}}{\frac{P_{sr}eB\Re}{L}W + \frac{4K_bT_nB}{R_L}}$$

$$BER = \frac{1}{2} erfc \sqrt{\frac{SNR}{8}} \quad (12)$$

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

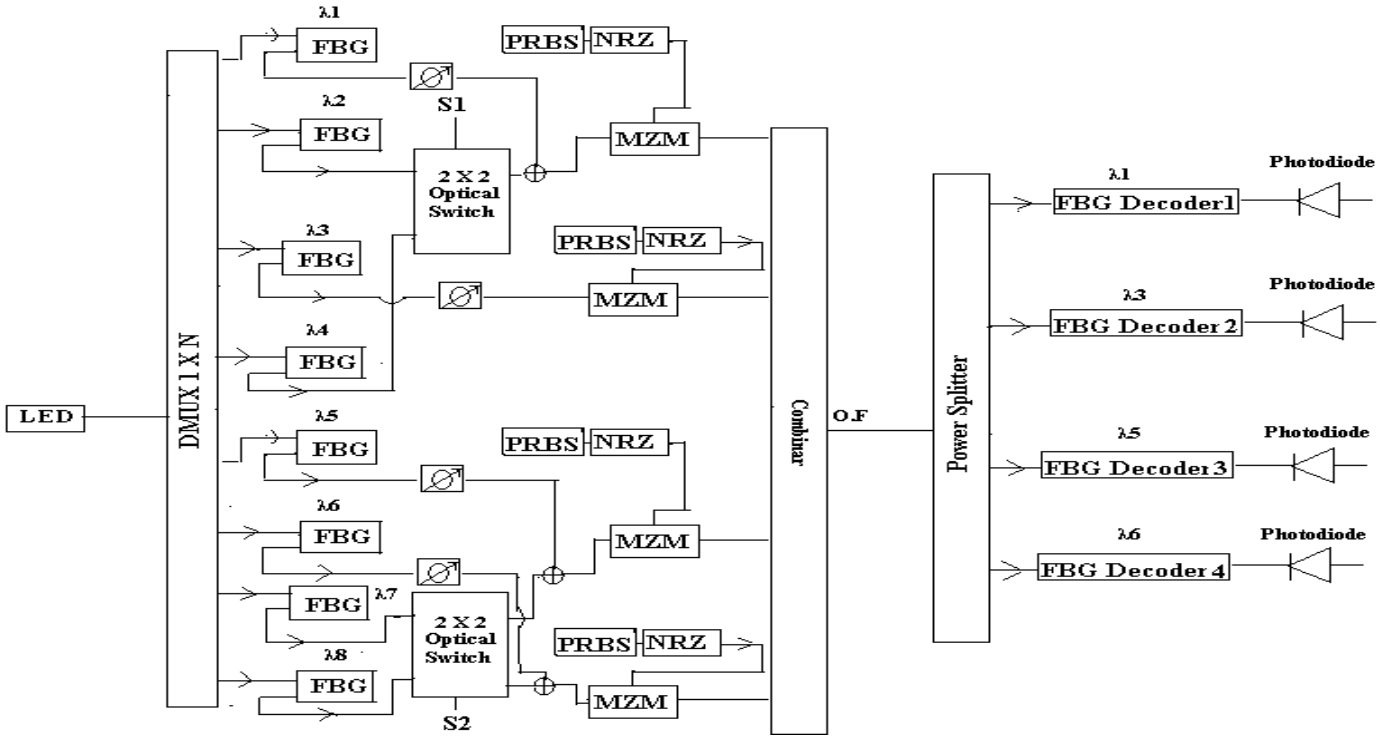


Fig1: Encoder and Decoder of Proposed Design

IV. RESULT AND ANALYSIS

Typical parameters used in the calculation as below:

Photo detector quantum efficiency ($\Re$)	0.6
Line-width broadband source (ΔV)	3.75 THz
Operating wavelength	1552 nm
Electrical bandwidth (B)	311 MHz
Data bit rate (R_b)	622 Mbps
Receiver noise temperature (T_n)	300 K
Receiver load resistor (R_L)	1030Ω

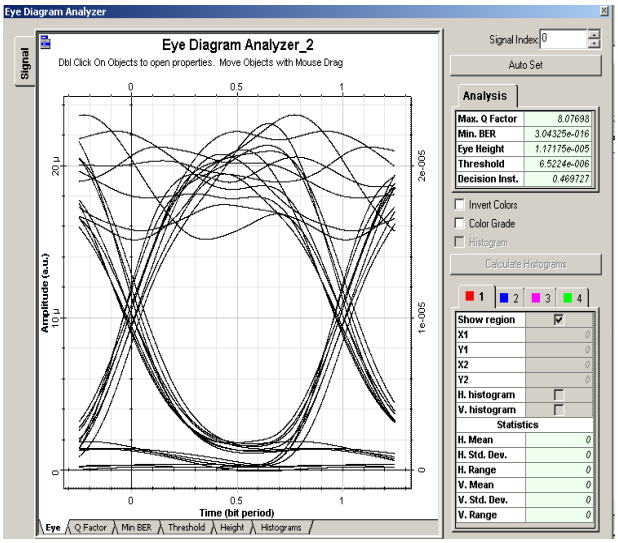


Fig. 2. Eye Diagram of 40 km for 6 users at 622Gbits/s

The block diagram of proposed scheme shown in fig 1 the simulation is done for 6 users with a weight of 2 .The width of each spectral chip kept 0.6 nm. The simulation is done in a practical environment in all, with all nonlinear effect is kept on. Simulation is performed for the 622Mbit/s for 40 km length of fiber and 1 Gbit/s for 40 km length of fiber with

ITU standard single mode fiber (SMF).All the attenuation (α= 0.25dB/km), Dispersion (18ps/nm) is maintained. Decoder side after decoding the signal, the signal covert to electrical by passing to the photo detector and 0.75 GHz low pass Bessel filter (LPF) The dark current value was 5 nA, and the thermal noise coefficient was 1.8×10^{-23} W/Hz for each of the photo-detectors. The performance of the system was characterized by referring to the BER and eye pattern.

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

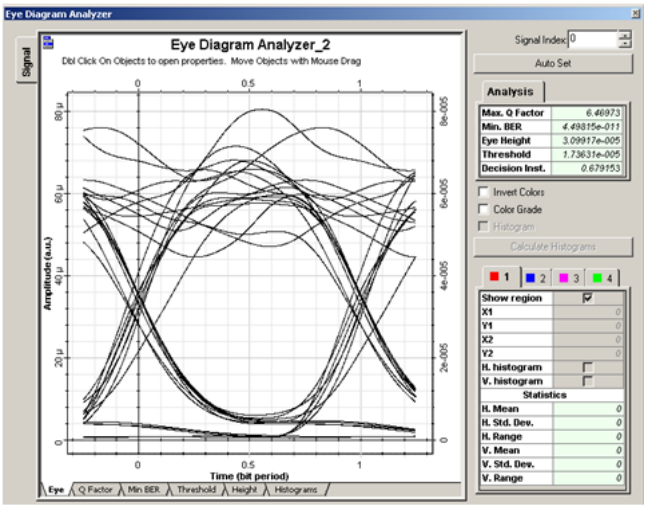


Fig. 3. Eye Diagram of 40km for 6 users at 1Gbits/s

A BER of the proposed Scheme shown that scheme gives better performance as the number user is increasing with comparatively lower bandwidth due to less number of weight width smaller spectral widths of each chip. Shown in fig the BER has provided more than minimum BER (10^{-9}). Fig 5 and 6 shows the probability of code detection for proposed technique and compare with the ZCC code [11] and MDW code. Where S represents the number switching code . The relation between power received at receiver end and BER shown in Figure7.

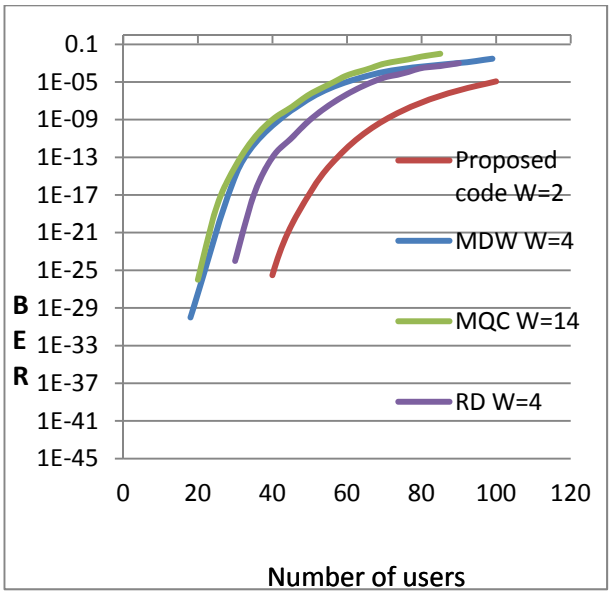


Fig. 4. BER performance vs. Number of users

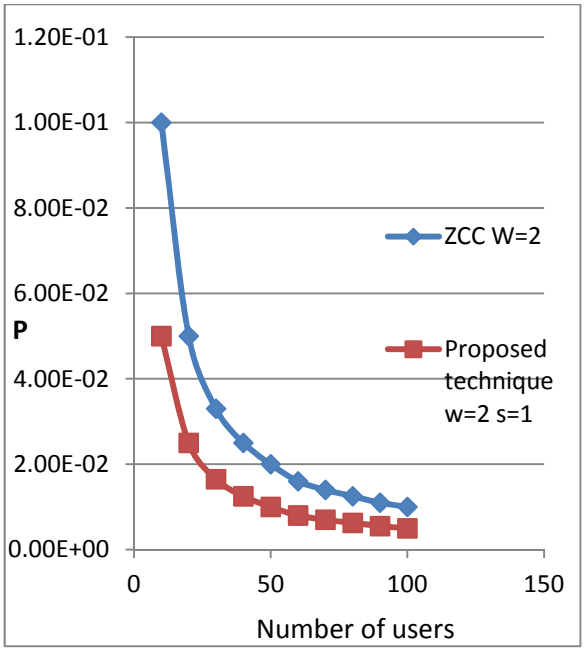


Fig. 5. Probability of code detetction Vs. Number of User for each pulse detection

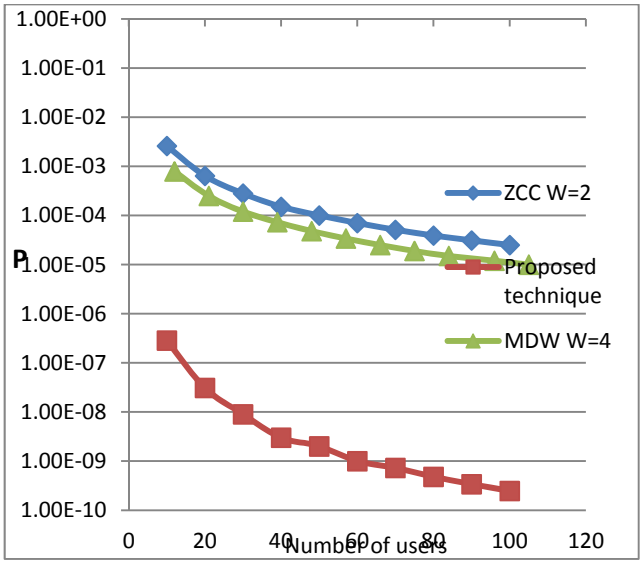


Fig. 6. Probability of code detection Vs. Number of User for whole code detection

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

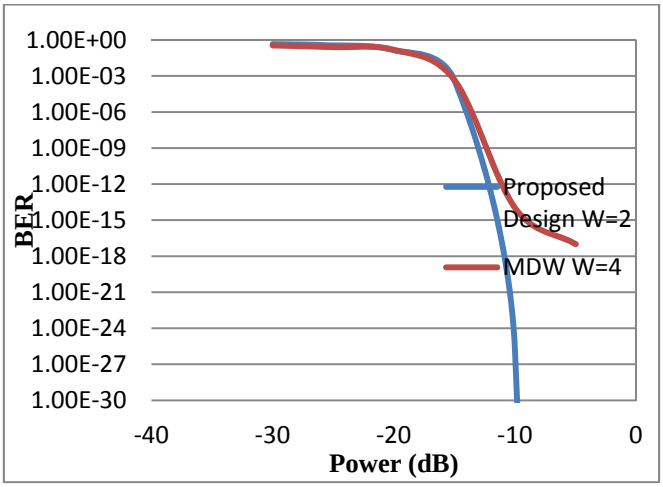


Fig. 7. Graph between the receive power (Psr) and BER

V. CONCLUSION

Security against the eavesdropper is enhanced using inter code switch between the user with different pattern from group to group. This technique reduces the code length of code design with better performance. Probability of information detection from single pulse of code or from whole code is better than the ZCC code. The performance also compare with existing code. Performance of proposed technique is analysis with optiwave7 simulation software and mathematical analysis.

REFERENCES

[1] J.A. Salehi, "Code division multiple access techniques in optical fiber network – Par I: Fundamental principles," IEEE Trans. Commun. 37 pp. 824–833, 1989.

[2] J.A. Salehi, C.A. Brackett, "Code division multiple access techniques in optical fiber network – Part II: System performance analysis," IEEE Trans. Commun. 37 pp. 834–842,1989.

[3] Shin-Pin Tseng, We Jingshown, "A new code family suitable for high-rate SAC OCDMA PONs applications," IEEE J. Select. Area Commun. Vol28 No.6 ,2010.

[4] S.A. Aljunid, M. Ismail, A.R. Ramil, "A new family of optical code sequence for spectral-amplitude-coding optical CDMA systems," IEEE Photon. Technol. Lett. 16 , pp.2383–2385,2004.

[5] E.D.J. Smith, R.J. Blaikie, D.P. Taylor, "Performance enhancement of spectral amplitude- coding optical CDMA using pulse-position modulation," IEEE Trans. Commun. 46 ,pp. 1176–1184,1998.

[6] Lei Xu, I. Glesk, V. Baby, P.R. Prucnal, "Multiple access interference (MAI) noise reduction in a 2D optical CDMA system using ultrafast optical thresholding," IEEE, vol. 2,Lasers and Electro-Optics Society, pp. 591–592,2004.

[7] Jen-Fa Huang, Chao-Chin Yang, "Reductions of multiple-access interference in Fiber-grating-based optical CDMA network," IEEE Trans. Commun. 50 vol.10 ,pp.1680–1687,2002.

[8] S.A. Aljunid, Z. Zan, S.B.A. Anas, M.K. Abdullah, "A New code for optical code division multiple access system," Malaysian Journal of Computer Science 17 (2) 30,2004.

[9] Hillal Adnan Fadhil, S.A. Aljunid, R.B. Ahmed ,"Per-formance of random diagonal code for OCDMA system using new spectral direct detection technique," J.Opt. Fiber Technol. 15, pp.283–289,2009.

[10] Zou Wei, H.M.H. Shalaby, H. Ghafouri-Shiraz, " Modified quadratic congruence code for fiber Bragg-grating-based spectral-amplitude-coding optical CDMA system," J. Lightw. Technol. 19 , pp. 1274–1281,2001.

[11]M.S. Anuar , S.A. Aljunid , N.M. Saad , S.M. Hamzah "New design of spectral amplitude coding in OCDMA with zero cross-correlation," Optics Communications 282 (14), 2659-2664 2009.

[12] Thomas H.Shake , Member, "Security Performance of Optical CDMA Against Eavesdropping," J. Lightw. Technol 23pp. 655-672,2005.

[13] Mohammad Ali Sedaghat, Ralf R. Müller, Farokh Marvasti "Performance analysis of asynchronous optical code division multiple access with spectral-amplitudecoding" *IET Commun.* Vol. 8, Iss. 6, pp. 956–963 2014

[14] Hossam M. H. Shalaby "Efficient Use of PPM in Spectral-Amplitude-Coding Optical CDMA Systems" *journal of Lightwave technology*, vol. 30, no. 22, 2012

http://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

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2015
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/