SPECIAL ISSUE

THE IIOAB JOURNAL

VOLUME 8 : NO 2 : JULY 2017 : ISSN 0976-3104

Institute of Integrative Omics and
Applied Biotechnology Journal

Dear Esteemed Readers, Authors, and Colleagues,

I hope this letter finds you in good health and high spirits. It is my distinct pleasure to address you as the Editor-in-Chief of Integrative Omics and Applied Biotechnology (IIOAB) Journal, a multidisciplinary scientific journal that has always placed a profound emphasis on nurturing the involvement of young scientists and championing the significance of an interdisciplinary approach.

At Integrative Omics and Applied Biotechnology (IIOAB) Journal, we firmly believe in the transformative power of science and innovation, and we recognize that it is the vigor and enthusiasm of young minds that often drive the most groundbreaking discoveries. We actively encourage students, early-career researchers, and scientists to submit their work and engage in meaningful discourse within the pages of our journal. We take pride in providing a platform for these emerging researchers to share their novel ideas and findings with the broader scientific community.

In today's rapidly evolving scientific landscape, it is increasingly evident that the challenges we face require a collaborative and interdisciplinary approach. The most complex problems demand a diverse set of perspectives and expertise. Integrative Omics and Applied Biotechnology (IIOAB) Journal has consistently promoted and celebrated this multidisciplinary ethos. We believe that by crossing traditional disciplinary boundaries, we can unlock new avenues for discovery, innovation, and progress. This philosophy has been at the heart of our journal's mission, and we remain dedicated to publishing research that exemplifies the power of interdisciplinary collaboration.

Our journal continues to serve as a hub for knowledge exchange, providing a platform for researchers from various fields to come together and share their insights, experiences, and research outcomes. The collaborative spirit within our community is truly inspiring, and I am immensely proud of the role that IIOAB journal plays in fostering such partnerships.

As we move forward, I encourage each and every one of you to continue supporting our mission. Whether you are a seasoned researcher, a young scientist embarking on your career, or a reader with a thirst for knowledge, your involvement in our journal is invaluable. By working together and embracing interdisciplinary perspectives, we can address the most pressing challenges facing humanity, from climate change and public health to technological advancements and social issues.

I would like to extend my gratitude to our authors, reviewers, editorial board members, and readers for their unwavering support. Your dedication is what makes IIOAB Journal the thriving scientific community it is today. Together, we will continue to explore the frontiers of knowledge and pioneer new approaches to solving the world's most complex problems.

Thank you for being a part of our journey, and for your commitment to advancing science through the pages of IIOAB Journal.

Yours sincerely,

Vasco Azevedo

Vasco Azevedo, Editor-in-Chief
Integrative Omics and Applied Biotechnology
(IIOAB) Journal

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

SECURE MOBILE BEACON BASED OBSTACLE AWARENESS IN WSN

S Velmurugan[1*] and E Logashanmugam[2]

[1]*Department of Electronics and Communication Engineering, St. Peter's University, Chennai, INDIA*
[2]*Department of Electronics and Communication Engineering, Sathyabama University, Chennai, INDIA*

ABSTRACT

Localization technology is mainly important to most applications of wireless sensor networks (WSNs). In WSN, Beacon node is send location information to the sensor node. But the obstacle presented environment cannot reach the location information. To overcome this problem, we propose Secure Mobile Beacon based Obstacle Awareness in Wireless Sensor Networks (SMBOA). In this scheme, the mobile Beacon is identified the obstacle present place and send the location information to the sensor nodes. The mobile beacon is verified based on the ID based signature scheme. The authorized mobile beacon sends original sensor location hence secure data transmission in WSN. The simulation result demonstrates that the SMBOA improve the throughput and diminish both the packet loss rate and delay. It also reduces the energy consumption and improves the lifetime of the network.

KEY WORDS
Localization, Security, Obstacle, Mobile Beacon, Wireless Sensor Networks.

INTRODUCTION

Localization is critical for several wireless sensor networking applications, including critical infrastructure protection, habitat monitoring, and target tracking. Additionally, location information employs an important part in bounding energy utilization in WSNs. Global Positioning System (GPS) is a usually used and specific method for sensor localization. Regrettably, the GPS method is neither expensive nor energy-efficient. The deployment capability of sensor nodes which are equipped with GPS may be reduced due to the increased size. Also, these GPS equipped sensors have inadequate applicability since GPS works only in an open field. Localization algorithms can handle with the difficulty where they are capable to deduce the location of sensors by using the information of the fixed positions of a few sensors. Usually, these small sizes of sensor nodes with recognized location information are called beacon. WSNs are usually applied for missions where human being process is not possible. Therefore, installing beacon nodes in a preset location is frequently infeasible. The beacon message broadcast the location is unworkable due to the obstacle. Also the precision of the localization raises the number of beacons, cost and energy utilization. Regarding to solve the above mentioned problem we using a mobile beacon can be working as an option solution to localize the whole network. Localization through the use of a mobile beacon is essentially more accurate.

The rest of the paper is ordered as follows: Section 2 summarizes the presented localization methods for mobile beacon and sensors. The secure Mobile Beacon based Obstacle Awareness method in Section 3. The simulation results are described in Section 4 followed by the conclusions in Section 5.

Related work

Exploiting Routing information [1] proposed solves the localization and data routing troubles in WSNs. The system is separated into 1-hop cluster and the localization is hierarchically performed inside each cluster, next among clusters. The benefit of the scheme is that no extra message is necessary to make localization, which is received by a ranging technique, exploiting network topology supplied by the routing algorithm. Clustering based Robust Localization (CluRoL) approach proposed to recognize the malicious anchors and reject them from the localization process. CluRoL explained the bound circle of an anchor with respect to an sensor as the circle whose center is at the anchor and whose radius is the estimation of the distance between the anchor and the sensor. A CluRoL approach helps each Sensor to localize itself perfectly, using a clustering mechanism that executes clustering of these proximal points. A Clustering Algorithm for Localization in WSNs (CFL) scheme [3] designing a localization algorithm based on clustering. This algorithm uses a combined weight function and tries to classify the sensor so that smallest number of clusters with highest number of nodes in each cluster could be achieved.

Cluster-Based Mechanism for Multiple Spoofing Attackers in WSN [4] to detecting spoofing attacks, determining the number of attackers when multiple adversaries masquerading as a same node identity; and localizing multiple adversaries. The spatial correlation of received signal strength (RSS) inherited from wireless nodes to detect the spoofing attacks. Cluster based mechanisms is developed to determine the number of attackers. Support Vector Machines method to improve the accuracy of determining the number of attackers. This algorithm provides strong evidence of high accuracy of localizing multiple adversaries. Cluster-based Localization Method [5] investigates the performance of cluster-based localization using received signal strength indicator (RSSI) used to develop a heterogeneous WSN consisting of inter connected body area networks, or clusters. This localization method is based on a cluster-based version of the Min- Max algorithm that eliminating the need to transmit a large number of localization request packets. This method improves the network's robustness and reliability and the safety of its users.

Received: 18 Nov 2016
Accepted: 28 Dec 2016
Published: 5 January 2017

***Corresponding Author**
Email:
velmuruganstpeters@gmail.com

Cluster based Iterative GPS-Free Localization (CIGL) algorithm [6] depends upon the distances between the sensor nodes and their neighbors obtained by measurements like Time of Arrival (TOA). It selects a subset of nodes to establish Local Coordinate Systems (LCSs) on the basis of clustering results. All the LCSs converge to form the global coordinate system and complete the nodes localization in succession. The successfully located nodes are chosen as new beacons to re-locate the remaining unknown nodes, namely expand localization coverage by iteration. This algorithm achieves localization accuracy and coverage, the communication range is small and the network deployment is sparse. Trusted and Secure Clustering [7] analyzed impact of signal-strength attacks while cluster communication on trust degree and level of security. This scheme provides adaptation of trustworthy and secure communication where information is ubiquitous.

Secure and efficient voting-on-grid clustering (VoGC) scheme [8] proposed to diminish the malicious beacon signal. VoGC method to reduce the computational cost, localization accuracy and identify malicious beacon signals. Mobile beacon-assisted localization algorithm based on network-density clustering (MBL(ndc)) [9] combines node clustering, incremental localization and mobile beacon assisting together. MBL(ndc) algorithm offers localization accuracy and reduce path length. Geometric constraint-based localization method disadvantage is increasing location error with enlarging the communication range [10].

Mobile-beacon assisted localization method [11] utilizes the geometric relationship of the perpendicular intersection to compute node positions. The drawback of this method is expensive. Localization scheme for WSNs using mobile anchors with directional antennas approach [12] for locating static sensor nodes by means of mobile beacon nodes equipped with four directional antennas. The method is efficient where the sensor nodes have no specific hardware requirements. Probabilistic localization algorithm [13] based on a mobile beacon utilizes TOA technique for ranging and uses Centroid formula based on distance information to estimate nodes location.

Received Signal Strength (RSS) based Localization for WSNs [14] investigated the possibility of SN localization by exploiting the inherent property of the WSN technology, particularly the RSS of the exchanged message. The RSS can be used for outdoor localization under well-defined topology constraints.

METHOD

In this paper, we propose Secure Mobile Beacon based Obstacle Awareness in WSNs. In this scheme, the obstacle is identified and secure mobile beacon node (MBN) based obtains location information of sensor node (SN). WSN consists of number of SNs which are randomly disseminated among obstacle in the predetermined environment. In a WSN, the SNs usually localize themselves with the help of BN that know their own locations. In this network, the few BNs are preset the network and it broadcast the sensor node position occasionally.

The obstacles appear in the network field and block to BN transmit the location information of the SNs. While the obstacle presents in the data transmission path, the fixed BN broadcast the sensor localization information is cannot reached to the SNs. Therefore, these SNs send notification message to the Base Station (BS). BS obtains this message and it send (Route Request) RREQ message to the authorized MBN. This RREQ message contain sensor ID, node position. The MBN accepts this RREQ and sending that place and act as the BN. The MBN is able to discover an unknown obstacle is moving or fixed where it occupies within the communication range. If the obstacle is immovable the MBN will work permanently otherwise it will work until that obstacle is change the position in the network. [Fig. 1] shows the architecture of the proposed Scheme.

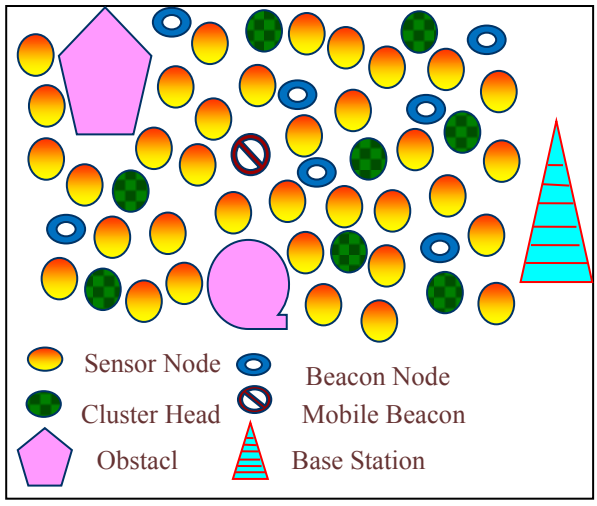


Fig. 1: Architecture of proposed scheme.

Mobile beacon node verification

The MBN received the message from BS and it going to the obstacle present place. Then the MBN send HELLO message to the obstacle near SNs. The SNs are received the HELLO message and checks this MBN is original or not. In this scheme, the MBN is verified based on ID based Signature method. It consists of three steps including Key setup, signature and verification.

• Key Setup: The BS computes a master key s and public parameters par for the Private Key Generation, and gives par to all sensor nodes.
A sensor node generates a private key R associated with the ID using the master key s.
• Signature: The mobile beacon generates a signature SIG based on the message M, time-stamp t and a signing` key.
• Verification: The SN checks the mobile beacon ID, M and SIG. If it match, the SN accept the MBN send the location information otherwise the SN reject the message.

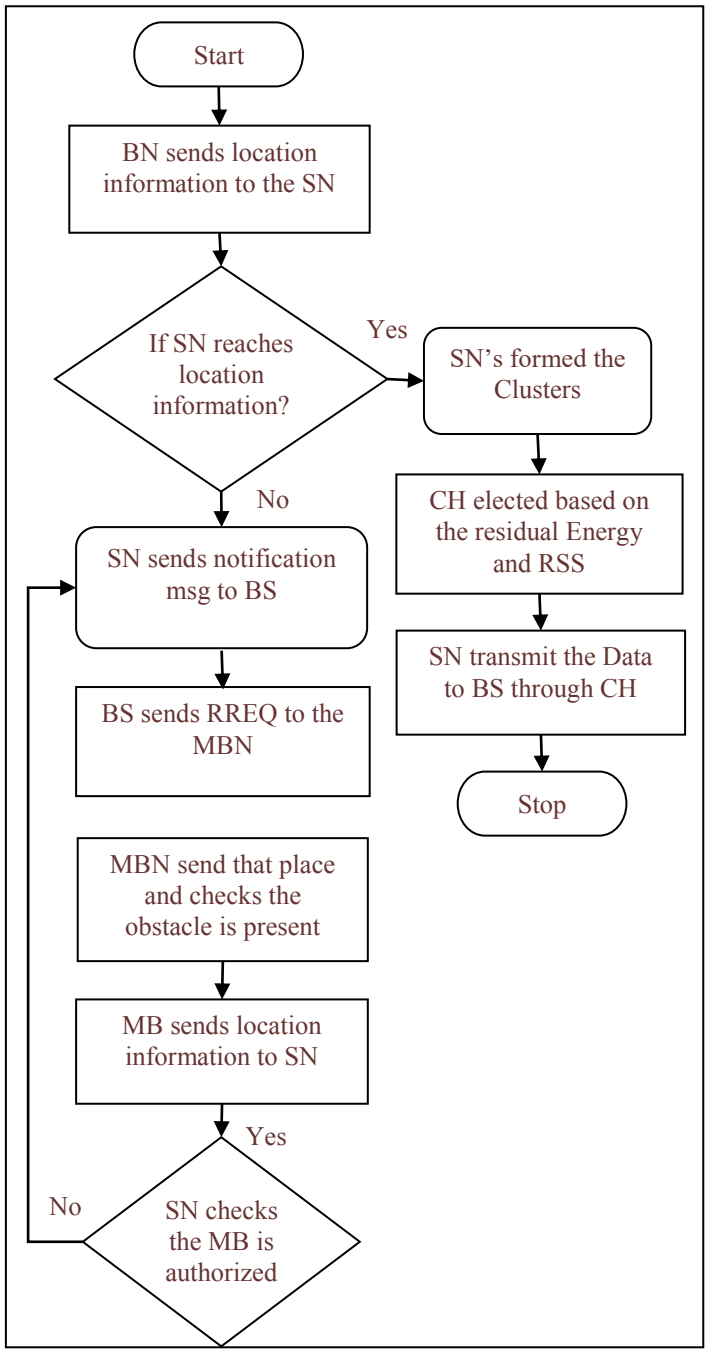


Fig. 2: Flowchart of SMBOA Scheme.

[Fig. 2] shows that the flowchart of the proposed scheme. The SNs are getting the location information from the BN or MBN and formed the clusters based on the coverage. Clustering is a standard approach for achieving efficient and scalable performance in WSNs. The sensor nodes are transmits the data through the cluster head. The Cluster Head (CH) is elected based on the highest residual energy and RSS. The SNs are transmits the data through the CH. This scheme used to reduce the energy consumption and improve the lifetime of the network.

Simulation analysis

The performance of the Secure Mobile Beacon based Obstacle Awareness in WSNs is examined by using the Network simulator (NS2). It is an open source programming language written in C++ and Object Oriented Tool Command Language. To estimate the proposed scheme we have assumed 65 sensor nodes, a network in an area of 1000x1500 m2. The parameters used for the simulation of the proposed scheme are tabulated in [Table 1]. Random waypoint mobility model is used to the sensor node movement. User Datagram Protocol (UDP) is used to node communication. We consider the packet delivery rate, packet loss rate, delay, residual energy are showing the efficiency of the proposed work.

Table 1: Simulation parameters of SMBOA

Parameter	Value
Number of nodes	63
Routing scheme	RSSL and SMBOA
Traffic model	Constant Bit Rate
Simulation Area	1000x1500 m^2
Channel	Wireless Channel
Transmission range	250m
Communication Protocol	UDP
Antenna	Omni Antenna

Packet delivery rate

Packet Delivery Rate (PDR) is the ratio of the total number of packets effectively delivered to the total packets sent. It is received from the equation (1) below.

$$PDR = \frac{Total\ Pkts\ Received}{Total\ Pkts\ Send} \quad (1)$$

The [Fig. 3] shows the PDR of the proposed scheme SMBOA is higher than the PDR of the existing method RSSL. The greater value of PDR means better performance of the protocol.

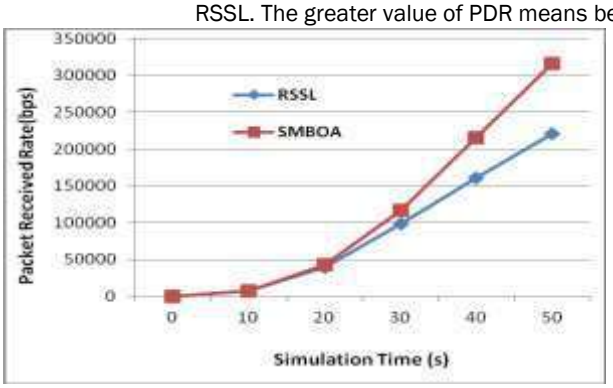


Fig. 3: Packet delivery rate of RSSL and SMBOA.

Packet loss rate

Packet Loss Rate (PLR) is the ratio of the packets lost to the total packets sent, estimated by the equation (2) below.

$$PLR = \frac{Total\ Pkts\ Dropped}{Total\ Pkts\ Send} \quad (2)$$

The PLR of the proposed scheme SMBOA is lower than the existing scheme RSSL in [Fig. 4]. Lower the PLR indicates the higher performance of the network.

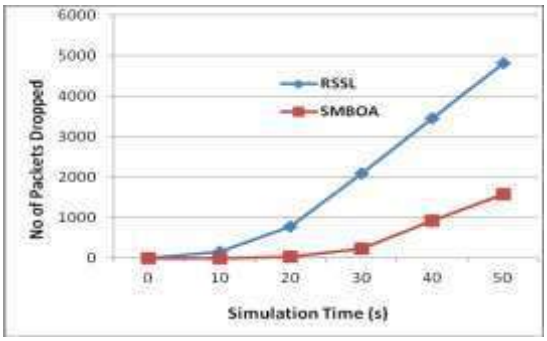


Fig. 4: Packet loss rate of RSSL and SMBOA.

Average delay

Delay is defined as the time difference between the current packets received and the previous packet received. Where n is the number of nodes.

$$Delay = \frac{\sum_{0}^{n} Pkt\ r\,e\,cvd\ time - Pkt\ send\ time}{n} \quad (3)$$

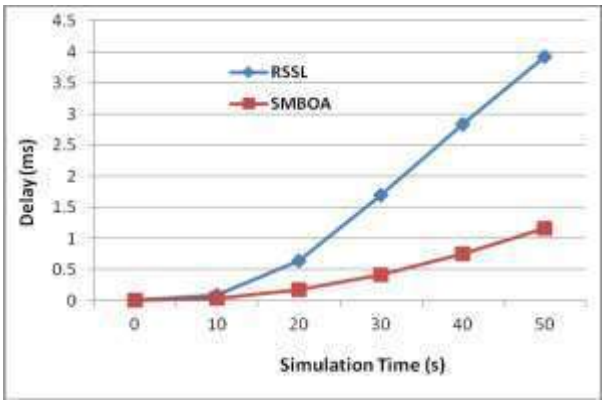


Fig. 5: Average delay of RSSL and SMBOA.

[[Fig. 5] demonstrates that the delay value is low for the proposed scheme SMBOA than the existing scheme RSSL. The minimum value of delay is improves the network performance.

Throughput

Throughput is defined as the rate at data is successfully transmitted for every packet sent.

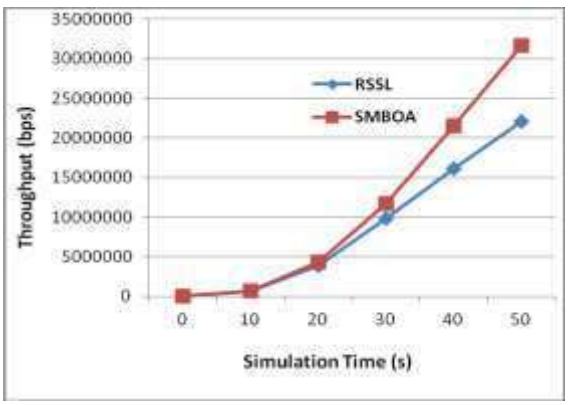


Fig. 6: Throughput of RSSL and SMBOA.

$$Throughput = \frac{\sum_0^n Pkts\ recvd\ (n) * Pkt\ size}{1000} \quad (4)$$

[Fig. 6] show that the proposed scheme SMBOA has greater average throughput when compared to the existing scheme RSSL.

Residual energy

The amount of energy remaining in a node at the current instance of time is called as residual energy. A measure of the residual energy gives the rate at which energy is consumed by the network operations.

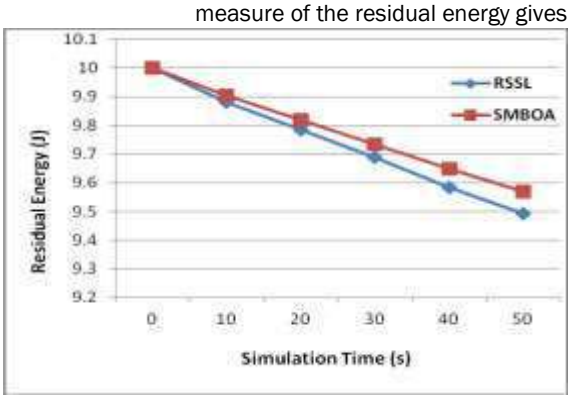


Fig. 7: Residual energy of RSSL and SMBOA.

[Fig. 7] shows that the residual energy of the network is better for the proposed scheme SMBOA when compared with the existing scheme RSSL.

CONCLUSION

Localization is important concept in WSN. In this paper we proposed Secure Mobile Beacon based Obstacle Awareness in WSNs. In SMBOA, the mobile beacon is sends the location information to the obstacle near sensor nodes. The ID based signature method is used to verify the mobile beacon. The clusters are formed based on the coverage. The Cluster Head is elected by Received Signal Strength and Residual Energy. The simulation result shows that the SMBOA improved the Packet delivery rate. It also reduces both the delay and energy consumption in the network.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Oliva G, Panzieri S, Pascucci F, Setola R. [2013] Exploiting routing information in Wireless Sensor Networks localization. In Network Science Workshop (NSW), 66-73.

[2] Misra S, Xue G. [2007] CluRoL: Clustering based robust localization in wireless sensor networks. In MILCOM 2007-IEEE Military Communications Conference, 1-7.

[3] Zainalie S, Yaghmaee MH. [2008] CFL: A clustering algorithm for localization in wireless sensor networks. In Telecommunications, 2008. IST 2008. International Symposium on, 435-439.

[4] Meena T, Nishanthy M, Kamalanaban E. [2014] Cluster-based mechanism for multiple spoofing attackers in WSN. In Information Communication and Embedded Systems (ICICES), 2014 International Conference on, 1-5.

[5] Zhong C, Eliasson J, Makitaavola H, Zhang F. [2010, 2014] A cluster-based localization method using RSSI for heterogeneous wireless sensor networks. In 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 1-6.

[6] Chen R, Zhong Z, Ni M. [2011] Cluster based iterative GPS-free localization for wireless sensor networks. In Vehicular Technology Conference (VTC Spring), 1-5.

[7] Gaur MS, Pant B. Impact of Signal-Strength on Trusted and Secure Clustering in Mobile Pervasive Environment. Procedia Computer Science, 57:178-188.

[8] Yang W, Zhu WT. [2010] Voting-on-grid clustering for secure localization in wireless sensor networks. In Communications (ICC), 2010 IEEE International Conference on, 1-5.

[9] Yang W, Zhu WT. [2010] Voting-on-grid clustering for secure localization in wireless sensor networks. In Communications (ICC), 2010 IEEE International Conference on, 1-5.

[10] Lee S, Kim E, Kim C, Kim K. [2009] Localization with a mobile beacon based on geometric constraints in wireless sensor networks. IEEE Transactions on Wireless Communications, 8(12): 5801-5805.

[11] Guo Z, Guo Y, Hong F, Jin Z, He Y, Feng Y, Liu Y. [2010] Perpendicular intersection: locating wireless sensors with mobile beacon. IEEE Transactions on Vehicular Technology,

59(7): 3501-3509.
[12] Ou CH. [2011] A localization scheme for wireless sensor networks using mobile anchors with directional antennas. IEEE Sensors Journal, 11(7): 1607-1616.
[13] Sun GL, Guo W. [2004] Comparison of distributed localization algorithms for sensor network with a mobile beacon. In Networking, Sensing and Control, 2004 IEEE International Conference on, 536-540.
[14] Stoyanova T, Kerasiotis F, Antonopoulos C, Papadopoulos G. [2014] RSS-based localization for wireless sensor networks in practice. In Communication Systems, Networks & Digital Signal Processing (CSNDSP), 2014 9th International Symposium on, 134-139.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

QUALITY OF SERVICE ON PERFORMANCE EVALUATION- A SURVEY

Deepa Mani[1*], Anand Mahendran[2]

[1]*School of Information Technology and Engineering, VIT University, Vellore-632014, Tamilnadu, INDIA*
[2]*School of Computing Science and Engineering, VIT University, Vellore-632014, Tamilnadu, INDIA*

ABSTRACT

Cloud environment is altogether different from conventional processing environment, and along with these, the execution of cloud performance is extra fundamentals. The development of information in the cloud is quick. Consequently, it requires that resources and framework accessible at removal must be similarly experienced. Infrastructure level implementation in cloud includes the execution of servers, system and capacity which go about as the absolute completeness for driving the whole cloud business. This paper aims at supporting investigation around the cloud computing and thereby giving an overview of the best in the class of QoS demonstrating methods reasonable for cloud frameworks. Our objective is to study overview of current and forthcoming investigation on QoS methods in cloud computing.

KEY WORDS
Dynamic power management; cost optimisation; Energy efficiency control; Reservation cluster;, systematic mapping

INTRODUCTION

Cloud computing has created in popularity starting late because of specialised and moderate favourable circumstances of the on interest limit organisation model. Various cloud executives are presently dynamically accessible, giving advertising, which includes Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) arrangements [4]. The development stack of the cloud has an additional standard in initiative data centres, where private and hybrid cloud structures are progressively included [2]. Despite the way that the cloud has essentially reworked the breaking point provisioning process, it represents to a couple of novel troubles in the region of Quality-of-Service (QoS) management [1]. QoS suggests the levels of execution, steadiness, and accessibility offered by an application and by the stage or framework that hosts it. QoS is essential for cloud customers, who foresee that supplier will pass on the advanced quality properties, and for cloud suppliers, who need to locate the right trade-offs between QoS levels and operational costs [3].

To accomplish indeterminate workload and to be exceedingly accessible for clients, anyplace at whatever point resource over-provisioning is a normal condition in a cloud system. Regardless, most power dependant facilities will unavoidably encounter the evil impacts of unmoving times or low use for some days or months. Since there, as a rule, have reduced activity realized by a method for unpredictable entries [5]. QoS properties have consistent consideration well before the initiation of cloud computing, implementation diversity and resource quarantine systems of cloud stages have altogether had QoS investigation, expectation, and certification [6]. Clustering is an essential and financially profitable podium for executing parallel applications that process vast measure of information with the hubs of a cluster through the interconnected system. Clustering is customarily utilised as a part of numerous information mining applications to gather together the measurably comparable information components [7].

Received: 19 November 2016
Accepted: 16 December 2016
Published: 5 January 2017

In case, all the above components have been satisfied we can say that cloud computing is working with the unequalled force. There is a chance that any of the factors comes up short or neglects to fulfill the client needs then we need to run further with other procedures of distributed computing. In the accompanying portions, we have examined about every one of the elements of Quality of Service of our best.

In this paper, we have explained a summary of the existing models that target to quantify and recover the power and energy consumption, workload, cost estimation and availability of data centres and cloud hosts. In section 2, the study on performance enhancement modeling is given. Section 3, describes performance evaluation modeling. This survey shelters performance enhancement modelling, power and energy consumption modelling, workload modelling, Cost optimisation modelling, Reservation cluster modelling, and the duality of performance modelling. Section 4 deals the different methods to achieve the low cost, energy consumption and high availability. Finally, we conclude in section 5.

MATERIALS AND METHODS

Performance enhancement modeling

***Corresponding Author**
Email: mdeepa@vit.ac.in
Tel.: +919789305010

Performance modelling aims to perform the evaluation or forecast of the reaction time not just central for little applications additionally for the vast applications that are running on an open cloud. The principle responsibility of the cloud server centers to offer the nature of administrations, paying little mind to the exuberant method for the cloud where the consignment changes were made [8]. To satisfy the quality need, demands encouraged on the public cloud ought to check for their presentation i.e. reaction time and preparing time so execution variables are inside the lenience boundary [9].

Power and energy consumption modelling

Power and energy consumption and estimation modelling involve gauging the vital utilisation of virtualized hardware. Resource dispute of numerous applications attempting to utilise hardware can bring about significant overheads [18]. Resource sharing can prompt a decrease in energy utilisation, just like the case in sharing a multicore CPU for instance. The shared memory of the CPU reserves the various centres which regularly prompts a performance increment, while less energy is being expended [11].

Workload modelling

Workload modelling demonstrates how to build a distinctive cloud workload pattern and prompts to more knowledgeable choices to achieve better resource management [12, 15]. Likewise, workload modelling in cloud computing empowers presentation scrutiny and reproduction, which conveys advantages to cloud suppliers and analysts as it permits. (i) the evaluation, through re-order, of supply organisation procedures allowing the change of cloud organisations' QoS.(ii) the assessment of these methodologies without sending and implementation of the requests in immoderate expansive scale circumstances; (iii) the imitation of sensible cloud situations using organised change, alteration, and redundancy [13,14].

Cost estimation modelling

Cost estimation modelling is requested turns for various sorts of resources over the period among which the in-house servers has displaced with virtual machine occurrences (hereon called the arranging time frame) [16]. In continuation to that, there are resources has been sorted like CPU, RAM, storage and network [17]. (i) Define the demand curve for CPU that is, the average number of runnable processes over the period. (ii) Define the demand curve for RAM that is memory usage of the server (iii) Define the network demand, i.e., some data dragged in and dragged out from the network. (iv) Define the storage demand, i.e., the amount of data deposited in the hard drives.

Availability modelling

Availability modelling in the cloud services is fundamental for upholding client certainty and staying away from income misfortunes because of Service Level Agreement infringement consequences [21]. Since the software and hardware segments of cloud foundations may have constrained dependability, the utilisation of unessential segments and numerous clusters might be mandatory to accomplish the anticipated level of reliability while also the vibrant increase in the gaining and the computational costs [20].

Performance Assessment modelling

Cloud computing assets must be perfect, superior and capable. High performance is one of the cloud preferences which must be acceptable for every administration. Higher execution of services and anything identified with cloud have an impact on clients and administration suppliers. Henceforth, execution assessment for cloud suppliers and clients is essential. There are numerous strategies for execution forecast and assessment; we utilise the accompanying techniques in our assessment [22].

- Assessment taking into account based on criteria and attributes
- Evaluation has given as re-enactment [23].

Variables affect on execution

These days, the tenure "performance" is an exemplary idea and incorporates more broad ideas, for example, reliability, energy effectiveness, scalability and so on. Because of the degree of cloud computing situations and the extensive quantity of undertakings and typical clients who are utilising cloud setting, numerous components can influence the execution of cloud computing and its capitals [10]. A portion of the essential elements deliberated in this paper are as per the following:

Safety

The effect of safety on cloud presentation may appear to be delicately weird; however, the effect of security on network infrastructure has demonstrated. For instance, dispersed denial-of-service (DDoS) bouts wide effect systems execution, and it will essentially diminish systems execution besides effective on reaction time also. Therefore, if this threat and any same threats undermine cloud domain, it will be a major anxiety toward clients and suppliers. Many researchers have been proposed in information security [23], insurance and access control to improve the security [24]. Cloud computing security gives a model to the customer centered information encryption for expanding the unwavering quality.

Data Integrity

Data integrity gives similarly adaptable, position-autonomous, low-cost stage for the customer. Data integrity comprises of two viewpoints. One is the productivity and safety in which to create people in general and private key plainly and proficient to do the secret key era. The bounds are determined, and it is productive since a lot of information encryption is done and keeps from attacks [24]. Since we scramble

information to keep from unauthorised clients; data integrity is kept up [25]. In information storage framework, customers store their data in the cloud for the accessibility of reports and the security must be guaranteed. One of the critical subjects is the Byzantine failure and has overcome in the distributed frameworks which may achieve the capacity issues. It comprises of survey and document appropriation, and homomorphic token is acquainted with doing the encryption [26].

Scalability

It is the capacity of the system to play out the predetermined functionalities which characterise its ability. Methodologies like horizontal scalability and vertical scalability have acquainted with enhancing the scalability of the system [27]. Cloud computing handles expanding requests. There are various sorts of scaling available they are vertical scaling is restricted by the way that we can just get the large size of the server, horizontal scaling manages the capacity to scale more extensive to manage traffic and corner to corner scaling [28]. Virtual machine versatility is constrained in the event of the TCP message workload contrasted with different threads. The condition of-workmanship engineering system permits various virtual machines to scale the length of memory breaking points. It is done until it achieves its cut-off points [29].

Performance

Execution change is the estimation of results of a specific procedure. An expansive measure of information can be isolated into lumps so that hacking of information can avoid without manipulating the whole information [30]. The elite presentation tests in two distinct fields, for example, supercomputing on a committed group and a group of virtual machines running in the cloud and different architectures has been proposed to enhance the execution rate [31].

System models

Power and energy consumption modeling

Measuring the energy utilisation of virtualized equipment is a long way. Resource dispute of various applications attempting to utilise hardware can prompt bring about noteworthy overheads [11].The resource sharing prompts a failure in energy utilisation, similar to the case in sharing a multicore CPU for instance. Getting to shared memory CPU stores prompts a performance increment, while less vitality has consumed. In this section, we make a refinement between usage based energy prediction models and performance monitoring counter-based (PMC) expectation energy models.

In usage-based models, direct regression is frequently connected to a particular arrangement of use hardware statistics and the deliberate energy utilisation. In PMC-based models, logged (virtualized) in which occasion hardware counters are utilised to frame a prediction model. It is less demanding to acquire OS-provided inputs, for example, CPU; disk and memory usage to attain the hardware event counters [33]. Be that as it may, these models can frequently not be widespread to all hardware setups, as proficiency of hardware segments contrast extraordinarily. In this manner, utilising relative PMC (rPMC) as is done in [32] can offer a more nonspecific and less blunder inclined model for evaluating the hardware utilisation for workloads.

Estimation modeling

The power utilisation of PCs is not about the work they fulfill [33]. The measure of unmoving nodes must be kept to a base in server farms since they consume an accommodating measure of energy while not performing any operations. The facts can confirm that a server needs to run, yet the just low execution is necessary. Dynamic voltage and frequency scaling (DVFS) assistance in diminishing the energy cost, permitting it to keep running in a lower power mode by downscaling the CPU voltage and frequencies. Virtual Machine Consolidation can likewise be viewed as a multi-target streamlining issue, taking the minimization of power utilisation and resource wastage as the destinations, as depicted in [34]. Resource wastage has seen as the whole of staying unused resource of a physical host.

The objective is to acquire a non-commanded set of arrangements, the Pareto set, utilising Ant Colony Optimization heuristics. Ant Colony Optimization (ACO) is a metaheuristic propelled by the perception of genuine ant colonies and based upon their aggregate scavenging performance [34]. The goal to do this, linear regression models for resource wastage and power utilisation framed, of which the outcomes utilised as the two destinations of the improvement algorithm. Based on the algorithm a requested set of VMs and physical hosts as info and figures the attractive quality. The likelihood of moving a VM to a particular host while ensuring that every host does not surpass its resource use limits. At the point when the Pareto ideal set has computed, VM consolidation can happen to utilise the subsequent VMs to hosts mapping.

Economic cloud federation

Cloud federations take into consideration relocating VMs between numerous data centres of various suppliers at different geological areas [11]. A cloud federation comprises of regularly slighter participating cloud suppliers that go for manageability, lessening carbon dioxide outflows, and bringing down power costs [35]. As little and medium cloud suppliers can frequently not resist with the huge cloud suppliers, for example, Amazon, Google, and Rack space, they shape a league that gives every member a chance to acquire an external resource from each other, while minimising the energy expenses and carbon-dioxide emissions of their data centres.

Power utilisation can minimise by taking the neighborhood power costs into thought. Each VM is allocated an energy spending plan for every economic time interim. In each economic time interim, the budget plans for all VMs in the following economic time interim are ascertained utilising an estimation of resource use for the VMs utilising linear regression.

Workload modeling
Workload categorization

The workload is the measure of handling that the PC has been given to do at a given time [42]. The workload comprises of some measure of utilisation programming running on the PC and typically some number of clients associated with and communicating with the PC's applications. The workload is characterised like computation, memory, networking, and storage [35].

Deployment environment: Factual characterizations of observational data are valuable in QoS [15]. Some of the QoS model parameters, e.g., network transfer capacity fluctuation, virtual machine (VM) start up times, begin failure probabilities used to estimate the practical qualities. Perceptions of performance variability have accounted for various sorts of virtual machine occurrences [35]. Different works describe the variability in VM start-up times , which is connected specifically with working operating system picture size.

Workload Implications

Regression Techniques. A typical workload derivation approach includes evaluating just the mean interest set by a given sort of requests on the resource [32]. The method depends on contrasting the performance measurements (e.g., throughput and usage of resource) anticipated execution model against estimations accumulated in a controlled test environment [4].

Cost estimation modelling

A cost capacity gives the details about the expenses of force utilisation, framework clog and server start-up. The impact of vitality productivity controls on reaction times, working modes and acquired expenses are all illustrated [33]. Our goals are to locate the ideal service rate and mode-exchanging confinement, to minimise the cost of a response time ensures under differing entry rates. A productive green control (EGC) algorithm is initially proposed for taking care of obliged optimisation issues and making costs/performance trade-offs in systems with various power sparing policies [5].

Amortization

It is critical to comprehend the commitment of IT base expenses. Consequently, amortisation limit is ascertained for servers and different amenities so that reasonable ascription of expenses for different IT resources (software/ hardware) can be realised [11]. The limit is mandatory to compute the month to month deterioration cost (amortisation expense) of every infrastructure. These things have introductory purchase cost, the expense of which is ascertained on the duration over which the speculation is amortised at the expected interest rate. Studies have uncovered that the expense of CPU storage and transfer speed twofold when the expenses have amortised over the duration of the substructure [35].

Cost of servers

It is acknowledged that each one of the servers has relative setups and servers mounted on racks. This statement is made to inspire the estimation of the cost of the server (without amortisation).Henceforth, the cost of the server has calculated with the combination of quantity of servers in a firm and the expense per server. The amortizable parameter for server determined in the past part will be used to choose the amortised server cost and amortizable Parameter for Server [34].

Availability modelling

There are some mechanisms which can help to protect against the failures, [32] have considered those mechanisms into groups like fault tolerance mechanisms, protective redundancy and overload protection.

Fault lenience mechanisms

Fault tolerance can be achieved, based on robustness and dependability of the system. It can classify into two types, i.e., proactive and reactive. The Proactive adaptation to non-critical failure arrangement is to

keep away from deficiency, blunders and failure by anticipating them. It will proactively displace the suspected part implies recognise the issue before it comes. Responsive fault tolerance to non-critical failure arrangements decreases the exertion of failure when the disappointment adequately happens. These can further portray by two sub-strategies like error taking care of and fault treatment. The main intentions of error handling are to expel the errors from the computational state. Fault treatment drives for keeping issues from being reactivated [31].

CONCLUSION AND FUTURE DIRECTION

In this paper first, we discussed facts that are involved in performance enhancement modeling and evaluation modeling. QoS approaches to receive a key part in the change of distributed computing to ensure that customers can trust cloud administrations. There has been a growing energy for QoS approaches in cloud computing among modern pros and analysts. The work comprehensively surveys the different performance enhancement modeling like power consumption modeling, cost estimation modeling, and availability modeling and so on. However, the technologies used in these modeling are somewhat difficult to analyze their corresponding QoS, from the service provider point of view. We have done detailed survey in workload and system modeling to QoS management.

CONFLICT OF INTEREST
There is no conflict of interest

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

COMPETING INTEREST
No competing interest

REFERENCES

[1] Armbrust M, Fox A, Griffith R, et al. [2010] A view of cloud computing. Commun ACM 53(4):50–58.

[2] Zhang Q, Cheng L, Boutaba R. [2010] Cloud computing: state-of-the-art and research challenges. J Internet Serv Appl 1(1):7–18.

[3] Akpan, Helen Anderson, B RebeccaJeya Vadhanam. A Survey on Quality of Service in Cloud Computing."*International Journal of Computer Trends and Technology (IJCTT) volume* 27: 58-63.

[4] Ardagna, Danilo, et al.[2014] Quality-of-service in cloud computing: modelling techniques and their applications. *Journal of Internet Services and Applications* 5.1 : 1.

[5] Chiang, Yi-Ju, Yen-Chieh Ouyang, Ching-Hsien Robert Hsu.[2015] An Efficient Green Control Algorithm in Cloud Computing for Cost Optimization.*IEEE Transactions on Cloud Computing* 3.2: 145-155.

[6] Petcu D, O Macariu G, Panica S, Craciun C. [2013] Portable cloud applications - from theory to practice. Future Generation Comput Syst 29(6):1417–1430.

[7] Malathy G, Rm Somasundaram.[2012] Performance enhancement in cloud computing using reservation cluster." *European Journal of Scientific Research, ISSN* : 394-401.

[8] Singh, Jitendra. "Study of response time in cloud computing." *International Journal of Information Engineering and Electronic Business* 6.5 (2014): 36.

[9] J Baker, C Bond, J Corbett, JJ Furman, A Khorlin, J Larson and, et. al.[2011] Megastore: Providing Scalable, Highly Available Storage for Interactive Services," Proceeding of Conference on Innovative Data Systems Research (CIDR), pp. 223-234.

[10] Grozev, Nikolay, and Rajkumar Buyya. [2015] "Performance modelling and simulation of three-tier applications in cloud and multi-cloud environments."*The Computer Journal* 58.1 :1-22.

[11] Warnaar, Martin.[2016] Cloud Energy Consumption Measurement and Reduction: an Overview of Methods.

[12] Magalhães Deborah, et al.[2015] Workload modeling for resource usage analysis and simulation in cloud computing. *Computers & Electrical Engineering* 47:69-81.

[13] Feitelson DG. Workload modeling for computer systems performance evaluation. Cambridge University Press; 2015. In press, available online at http://www.cs.huji.ac.il/~feit/wlmod/wlmod.pdf.

[14] Moreno I, Garraghan P, Townend P, Xu J. An approach for characterizing workloads in Google cloud to derive realistic resource utilization models. In: Proceedings of 7th international symposium on service oriented system engineering (SOSE). IEEE; 2013. p. 49–60.

[15] Chen Y, Ganapathi AS, Griffith R, Katz RH. Towards understanding cloud performance tradeoffs using statistical workload analysis and replay. Technical Report. University of California at Berkeley; 2010. URL: http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-81.html.

[16] Truong, Hong-Linh, Schahram Dustdar.[2010] Composable cost estimation and monitoring for computational applications in cloud computing environments.Procedia Computer Science 1.1 : 2175-2184.

[17] Singer, Georg, et al.[2010] "Towards a model for cloud computing cost estimation with reserved instances." *Proc. of 2nd Int. ICST Conf. on Cloud Computing, CloudComp 2010*..

[18] Vilaplana, Jordi, Francesc Solsona, and Ivan Teixidó.[2015] A performance model for scalable cloud computing. *13th Australasian Symposium on Parallel and Distributed Computing (AusPDC 2015), ACS*. Vol. 163.

[19] P.Latchoumy, P.Sheik AbdulKhader. [2011] Survey on fault tolerance in grid computing" IJCSI International Journal of Computer Science Issues, 2(4)

[20] Dantas, Jamilson, et al.[2015] Eucalyptus-based private clouds: availability modeling and comparison to the cost of a public cloud." *Computing* 97(11) :1121-1140.

[21] Nguyen, Tuan Anh, Dong Seong Kim, Jong Sou Park. [2016] Availability modeling and analysis of a data center for disaster tolerance." *Future Generation Computer Systems* 56: 27-50.

[22] AYMAN G. FAYOUMI, [2011] PERFORMANCE EVALUATION OF A CLOUD BASED LOAD BALANCER SEVERING PARETO TRAFFIC, Journal of Theoretical and Applied Information Technology, 32(1).

[23] Lar, S-U., Xiaofeng Liao, and Syed Ali Abbas. "Cloud computing privacy & security global issues, challenges, & mechanisms." Communications and Networking in

China (CHINACOM), 2011 6th International ICST Conference on. IEEE, 2011.
[24] Kulkarni, Gurudatt, et al. [2012] A security aspects in cloud computing." Software Engineering and Service Science (ICSESS), 2012 IEEE 3rd International Conference on. IEEE.
[25] Chalse, Rajkumar, Ashwin Selokar, Arun Katara.[2013] A New Technique of Data Integrity for Analysis of the Cloud Computing Security. Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on. IEEE,
[26] Ramaiah Y, Govinda G, Vijaya Kumari.[2013] Complete Privacy Preserving Auditing for Data Integrity in Cloud Computing. Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on. IEEE.
[27] Lee Jae Yoo, Soo Dong Kim. [2010.]oftware approaches to assuring high scalability in cloud computing." e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on. IEEE,
[28] Hassan, Shoaib, and Farooque Azam[2014]. Analysis of Cloud Computing Performance, Scalability, Availability, & Security." Information Science and Applications (ICISA), 2014 International Conference on. IEEE,
[29] Jamal, Muhammad Hasan, et al.[2009] Virtual machine scalability on multi-core processors based servers for cloud computing workloads." Networking, Architecture, and Storage, 2009. NAS 2009. IEEE International Conference on. IEEE,
[30] Behal, Veerawali, and Anil Kumar. "Cloud computing: Performance analysis of load balancing algorithms in cloud heterogeneous environment." Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-. IEEE, 2014.
[31] Keville, Kurt L, et al. 2012] Towards Fault-Tolerant Energy-Efficient High Performance Computing in the Cloud.", Cluster Computing (CLUSTER), 2012 IEEE International Conference on. IEEE,.
[32] P Xiaoa, Z Hub, D Liua, G. Yana, and X. Qua.[2013.] Virtual machine power measuring technique with bounded error in cloud environments. Journal of Network and Computer Applications, 36(2):818{828,
[33] A Kansal, F Zhao, J Liu, N Kothari, A.A Bhattacharya. [2010] Virtual machine power metering and provisioning. In Proceedings of the 1st ACM symposium on Cloud computing, pages 39{50).
[34] Y. Gao, H. Guan, Z. Qi, Y. Houb, L Liuc. A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. Journal of Computer and System Sciences, 79(8):1230{1242, 2013.
[35] M Giacobbe, A Celesti, M. Fazio,
[36] M Villari, A. Pulia.[2015]to. An approach to reduce carbon dioxide emissions through virtual machine migrations in a sustainable cloud federation. In Sustainable Internet and ICT for Sustainability (SustainIT), pages 1(4).

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

ILLUSTRATION OF CODE CHANGE IMPACT ANALYSIS TOOL TO PREDICT THE SOFTWARE PROGRAM BEHAVIOR

Ashwini J Patil*, Netra Patil
Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Pune, INDIA

ABSTRACT

In early days, every Software applications are progressively huge as well as complicated, so it required enlarging in terms of prediction and manages the consequence of software application alteration. Alteration of Software application is a procedure of choosing the alteration to promote, as per the software project condition such as agenda and expenditure, which change to permit as well as which change to avoid. Therefore, this procedure discovers the modifications, source, classifies important project verdict positions, and creates project positions and tasks. This study presents a tool called code change impact analysis that helps to distinguish the classes of necessity modifications which has analogous impact stages. With prediction the impact of software code that necessity alterations may have the outcomes of building a necessity alteration can be evaluated to further necessity alterations with respect to the predicted attempt to execute the alteration. With this consideration the fact which we have targeted is 'Impact Analysis' tool that helps software manufacturing industry to predict difficulties from the source code that may arise after carrying out some development. This study has used two different algorithms which help to predict the alteration in the source code. The first classic data mining algorithm help to predict the problems and clustering algorithm helps to assemble these predicted problems. This impact analysis tool facilitates to decrease preservation effort as well as the hazard of costly alters. With this Impact analysis data we can do changes in planning, in creations, and accepting assured types of software, plus tracing throughout the effects of alterations. Also decrease the hazards of boarding on an expensive alter since the rate of unpredicted problems usually enlarges among the delay of their detection.

KEY WORDS
Impact analysis, necessity modification, predict problem

INTRODUCTION

A classification of the promising expenditure of a modification, or a calculating what desires to be customized which helps to achieve a modification is called as change impact analysis. Estimation of hazard is also linked among the job of impact analysis technique. The incentive in the impact analysis technique is nothing but to recognize effort goods which may be exaggerated via a modification. To effectively management of modification in the software application expansion, different procedures are requisite to offer the recorded data regarding modifications made in the software application, like how will the modification crash an expansion plan plus what modifications will have the maximum crash on the software application. With recorded information regarding modification, suitable preparation can be done via software project organization for applying or suspending modifications.

In recent times, the rising appreciation of the system necessities the engineering creates an expanding understanding on necessities traceability plus change impact analysis technique, that also demands a huge claim for a organized proceed in promising software application systems that running traceability associations plus necessities modifications in the repeated manner. A target determined necessities traceability method is manufacture plus hold necessities modifications close to three scopes like spread out software application plus hold necessities that are locate on the target determined utilize case method as well as to set up plus hold the traceability correlation amid a plan constitution matrix to achieve the traceability tree, plus to calculate necessities modification impacts through the isolating of plan constitution matrix addicted to amounts to afford as a core for the controlling exercise case positions [1]. There is different dissimilar software expansion lifecycle phase are available to Expand software module in that software maintenance plus exploitation is a chief expensive phase, among specific assessment since 70 to 90 percent of the complete rate. There are several numbers of software implementation developers generally consent amid the constructing software application modifications among no visibility addicted to their reasons which are express to contemptible to try estimates, delay in the discharge the policies, trouble in software module blueprint, unreliable software applications, plus precipitate departure of the software application .Software alteration impact analysis scheme offers significant organize in perceptive plus executing transform in the system application while it suggests a inclusive assessment of the rate of modifications in software application . Impact analysis scheme suggests transparencies in the feasible outcomes of modifications earlier to the actual modifications that are accomplished. The potential to distinguish the change impact or a practicable consequence that will extensively support a maintainer to corroborate right operate to obtain amid significance to modification resolution, strategies, rate as well as supply opinion [2]. Ensuing impact analysis and its software module traceability help convey the rate or ripple-effects of a designed modification diagonally unrelated step of software application modules. Experiment in the necessity traceability in its competence to integrate the huge point among the small point software application modules to aid engage the necessities, analysis cases, intend plus program source code [3]. There is multiple dissimilar research learning has been executed on the free source software application called SoMoX1. SoMoX is called as a software metric evaluation device that capable to re-engineering a module institute from the software's achievement that contains a set of Eclipse plug-ins through inaccurate 10,000 lines of Java program source code. SoMoX tool make a use of an Abstract Syntax Tree model as an input data model that helps to concludes method locates on hierarchical clustering of a vital methods acknowledged since classes plus interfaces [4]. In the software system maintenance it occupies an impact analysis tool as well as the communication of

Received: 23 Nov 2016
Accepted: 30 Dec 2016
Published: 15 Feb 2017

***Corresponding Author**
Email:
ashwinip684@gmail.com
Tel.: 9921316224

modification. WHAT-IF Tool is integrated into each software module expansion impression neither it will be accomplished among a huge software modules although the effects of the laboratory examination are extremely heartening [5]. There are different commands of agile development on the approval modification that will straight to a set of modifications during improvement. There are different dissimilar modification to efficient necessities, innovative limitation utility preferred, modifications to the software's module location, modifications to the system hardware, a narrative satellite is recognized plus modifications appropriate to unpredictability create during checking [6]. Different dissimilar Data mining algorithms have been recently associated to the software application warehouses to support on the software maintenance of budding software modules [7]. In the proposed system we have implemented a code change impact analysis tool which helps to determine classes of necessities modifications that have associated impact stages. Via predicting the impact, necessity modifications can have the effects of building a necessity modification can be estimated to further necessity modifications among approval to the predicted challenge to execute the modification. This information records can be useful in an evaluation of a method that chooses which changes can be implemented in plan limitation. It uses a standard data mining algorithm which helps to predict the problems plus clustering algorithm to group the predicted problems.

This research paper gives special features:

- It evaluates the effort of modifications.
- Cut the efforts of software maintenance rate.
- Reduction of hazard in the expensive modification.
- Allow managing board to build tradeoffs amid unusual Modifications.

This research paper is structured as follows: we have survey the prior unusual impact analysis methods for a software application system in section 3. Proposed new method towards source code change impact analysis tool in section 4. We have depicted a conclusion and future work in section 5.

Background and motivation

The motivation behind this proposed system is by the current requirement increasing the capability of software application progre. Multiple special previous attempts have focused on impact analysis technique for software application maintenance modifications to ended software component amid traceability. Software application system maintenance desires a tool for an impact analysis as well as the communication of a modification. Modification analysis is a critical phase to software application system maintenance, also is a total of numerous procedures via alter achievement. Change impact analysis can support improving programmer competence in multiple ways. Accurateness in change impact analysis that assurance the correctness as well as entirety of the software expansion growth. New examination on impact analysis is stands on the source code analysis.

Change impact analysis schemes necessitate being valuable initially at the structural design phase to control component reliance among no being deprived on coding scheme. Therefore we have centered on change impact analysis by eliminating the dependences between dissimilar classes from program source code.

Existing system

Malcom Gethers et al. has presented a method to accomplish impact analysis from a particular alter order to source code program [8]. Definite a textual alter order, an exacting instant of source code program fragment that are indexed via Latent Semantic Indexing, is used to utilize near the impact set. It should further contextual data can be offered, this method builds the best-fit combination to produce a better impact set. Contextual data encompasses the performance chart plus a primary program source code part entity definite for alter. Combinations of data recovery, active investigation, plus data mining of pattern program source code segment allocates are considered. The research theory is a combinations that aid counter the accuracy or again underperformance of mortal schemes as well as advancement of the mostly exactness. The three schemes set that isolated from added associated justifications. Automation among the capable employment of two key assets of developer data, that is repeatedly unobserved in impact analysis scheme at the alter order stage, that is skilled. To confirm proceed; they have carried out a tentative opinion on four open resource software applications. A standard including of a number of shield troubles, like as aspect requests plus virus join, and their associated program source code alters was acknowledged via static examination of these software systems and their alter documentation. Their conclusion spot to that there are mix produced from the amplified developer contextual data that prove a statistically key growth over the objective progresses.

Suhaimi Ibrahim et al. have covered a software traceability procedure which helps to retain change impact analysis of the object oriented software system [9]. Traceability procedure is an important part which can be experiential in its potential to integrate the high rank among the low rank software application representations that can engage the supplies, blueprint, assessment cases, plus code. This procedure authorizes during association between mechanisms at one rank to other mechanism at every ranks. It retains the top down plus bottom up traceability in respond to tracing for the ripple-effects. They have also extended a software model called Catia that hold C++ software, which is valuable to a study assessment of an embedded system that also shown in the outcomes.

S. M. Ghosh et al. In this research they have projected a technique called impact analyses [10] that make a use of sequential modification records for performable plus non performable files in a software application to distinguish as well as prioritize probably inclined region of a software system modification that support on vulnerabilities. This Change Impact Analysis procedure detains the most recent data on the knowledge and ability of determining that what software application sections manipulate all further. They have also shown a scheme for impact analysis enhanced. They have described a structure for the field, plus underline deliberation on very important outcomes. To distinguish key impact analysis clarifications, plans plus express the significant plans to present a compact aware for trying impact analysis difficulties mainly in ERP.

Neha Rungta et al. have shown iDiSE, an expansion to DiSE which helps to accomplish an interprocedural examination [11]. iDiSE join static plus dynamic describing environment data to resourcefully generate mandatory program behaviors diagonally describing environments. Forced program behaviors information is supportive with respective to the analyzing, confirmation, plus repairing of increasing source codes. Implementation of the case study is an iDiSE algorithm which helps to illustrate its efficiency at figuring impacted source program actions. In this they have also describe a new justifications of impacted exposure metrics which help to estimate the experimenting critical to inspection growing source programs. Then they have explained how the plans of impacted exposure which can be useful to categorize procedures like as DiSE plus iDiSE organize to retain regression testing interconnected jobs. It Also discuss how DiSE as well as iDiSE can be controlled for correcting; determining the center cause of faults launched through alterations fulfilled to the source code program. In the experiential study they have shown that the configurations of DiSE as well as iDiSE can be useful to retain dissimilar software maintenance works.

Mark Sherriff et al. have tried to verify the impact of a narrative system modification via investigating software source program alter records throughout exact significance decomposition [12]. In this scheme, it generates factions of files that conventionally are probable to alter concurrently to the address faults plus discontents begin in the source code base. They accomplish a post hoc case assessment via this scheme on five open source software applications. They have also created scheme that was winning in distinguishing impacted files in a software system from a recognized alter when the developers are tending to construct small, besieged changes to the source software system often. They more estimated their scheme different two added impact analysis schemes that is Path Impact plus Coverage Impact also they create that their scheme offered correspondent outcomes, but also distinguishing non-source code files which can be impacted via the alter.

Bixin Li et al. have described narrative process called a Software change impact analysis which is a procedure that discovers the roots of a modification, or resembling that what needs to be adapted to accomplish a alter [12]. Whereas the 1980s, there are so many assessments done on this procedure, principally for code-based alter impact analysis schemes. This research paper has resists to accomplish this breathing space. Also 30 research papers which recommend tentative evaluation on 23 source code-based alter impact analysis schemes are acknowledged. Then, information was grouped next to four learn questions. This research study proves a comparative outline that includes seven lands, that helps to discriminate the alter impact analysis schemes, plus distinguish key functions of alter impact analysis schemes in software system protection. In computation, they necessitate for more examine is also presented in the next two measurements: approximating open alter impact analysis schemes plus recommending latest alter impact analysis schemes in the proposed structure, growing further launched tools that helps to sustain alter impact analysis, estimating current alter impact analysis schemes empirically amid an incorporated metrics plus usual principles, as well as concerning the alter impact analysis further widely plus profitably in a software system maintenance phase.

Problem definition

The objective of this proposed system is to implement a predictive impact analysis tool to facilitate identifies classes of necessities modifications which have related impact stages. Via predicting the impact to necessity modifications can have, the effects of assembling a necessity alteration can be evaluated to other necessity alterations through the predicted effort to execute the modification. This information is used as a condition in a method that decides which changes can be executed contained by plan constraints.

Proposed system

In this proposed system, a code change impact analysis tool that facilitates to recognize classes of necessities alterations which has analogous impact stages.

Here, Impact analysis is a method to predict as well as confirm the elements of a software application which may be influenced by alterations to the software system. With the prediction of the impact, the necessity alterations may have the causes of building a necessity alter that can be evaluated to other necessity alterations with respect to the predicted attempt to execute the alteration. This alteration data can be used in the process that chooses which alterations can be executed in schedule limitations. It uses two algorithm which helps to predict the problems as well as group these predicted problems.

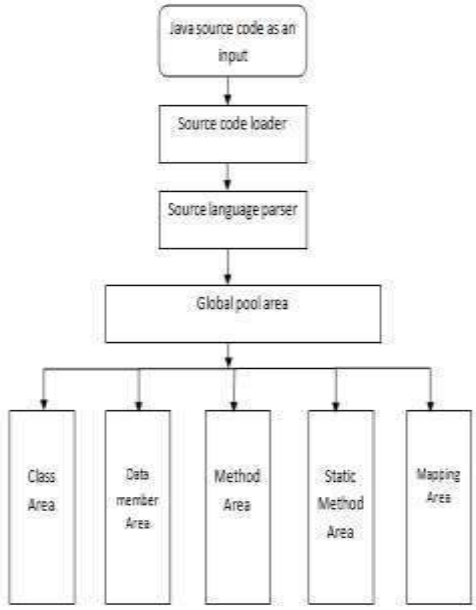


Fig. 1: System workflow.

This proposed tool uses java source code is as an input [Fig.1]. In this source code loader loads the source code as well as the source language parser is responsible for parsing the data. Therefore the results are in the form classification of the data. In this change set illustrates elements of the software application that are to be modified and Impact set illustrates elements of the software application which are influenced by the alterations. We recognize the region in which we want to make the alteration. Once discovering the ripple cause which may arise due to this alteration. Then we will come to a conclusion whether this alteration is sensible or not [2]. Through identifying possible impacts via making an alteration, we extensively decline the hazards of inflowing on a costly alteration because the price of unexpected difficulties generally improve during the interruption of their discovering. It builds the feasible outcomes of alterations that are capable to be seen by the alterations are implemented to create it unforced to implement alterations more properly and distinguish the charge or ripple effects of planned software alterations during progress as well as maintenance phase.

Proposed system has two different models are called as Work Information Model and Requirement Information Model, Where the work goods in familiar are those requirements which are altered. Work Information Model shows the fundamentals of a software expansion scheme which are to be utilized for impact examination. In the graph based model, work goods are assumed as nodes, plus the traces are characterized intended for edges among the nodes. Utilities are classified for every attribute to relate the work good or trace characteristic among the allocated stage. Work Information Model is defined as:

$$W\,IM = hNodes, T\,races, i, c, e, pi$$

Where Nodes shows the list of work goods, T races shows the list of edges, I for influence function, c for complexity function, e for effort function and p for phase cost function. Requirement Information Model contains information on the requirement alters which are the center for the examination. It is also graph based model through requirement alters showed as nodes as well as intended for edges characterize traces since requirement alters to distressed requirements. A utility links every edge through the allocated cruelty of the alter.

Requirement Information Model defines as:

$$RIM = hChangeSet, RequirementNodes, ChangeT\,races, ii$$

Where, ChangeSet is the list of requirement alterss, Requirement Nodes shows the list of work goods, ChangeT races is for list of edges, i is for function.

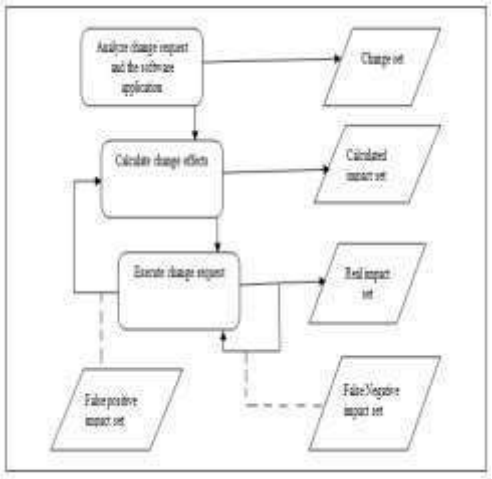


Fig. 2: System process flow.

Proposed system [Fig. 2] is divided into following modules are as follows:

Source code validation and parsing

In this it report validation error found in java source file. Every errors will located in the present window's location record list that is help location-list plus the equivalent lines in the source code file will be noticeable through Vim's :sign functionality among the '>>' indicators in the left edge. Automatic validation for the java source code files can be halt through the g:EclimJavaValidate variable. If we select to halt automatic validation, still we can utilize the :Validate control to physically confirm the present file.

Parsing or syntactic analysis is the procedure of evaluating a string of signs, whichever in normal language or in computer languages, by the policies of a formal grammar. Therefore it validates the input and removes the Meta data information from the specified input class.

Meta data generation

The Meta Data in java source file is a specified set of suggestive and structured data on a cluster of computer data. It is essentially used in the java programming language. This type of data contract with the structure, permission, storage space and exchange of Meta Data in the programming language. This module is essentially considered to incorporate the call that created .xml files into methodical manner since on demand of Matrix creation algorithm which we can simply offer the mandatory aspects.

Clustering algorithm

Clustering algorithm is a main task of grouping a set of objects in such a manner that objects in the similar group are additional similar to every other than to those in other groups. It is a key mission of examining data mining, as well as a general method for statistical data analysis

Report generation

It will generate all the graphs which will illustrate the all stages of the clusters. In this Google chart library is used to produce the graphs [Fig. 3].

Fig. 3: The java source code as an input and select class, method to process.

When saving a java source file that resides in a project, eclim will update that source file in Eclipse and will report any validation errors found. Any errors will be placed in the current window's location list (:help location-list) and the corresponding lines in the source file will be marked via vim's :sign functionality with '>>' markers in the left margin.

Aotomatic validation of java source files can be disabled via the g:EclimJavaValidate variable. If you choose to disable automatic validation, you can still use the :validate command to manually validate the current file.

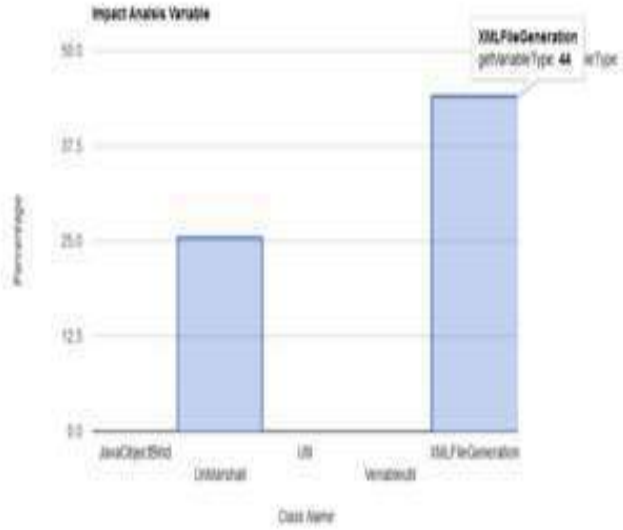


Fig. 4: The impact of proposal change on various classes.

The Fig. 4 shows the impact of proposal change on various classes. On X-axis we have those classes which will get affected due to the proposal change. On Y-axis we have the percentage of how much the particular class will be get affected due this change.

For example due to change in get Variable method, the class XLFileGeneration will get affected 44 percent. Unmarshall class will get affected by 26 percentage and so on.

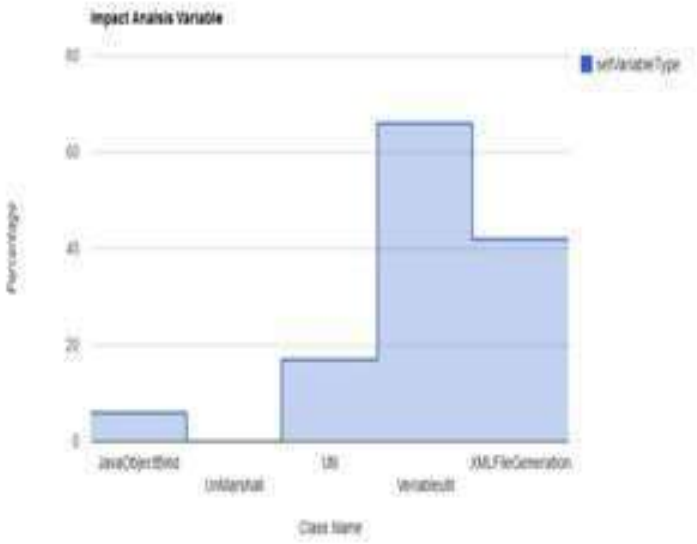


Fig. 5: Predict the percentage of ripple effects.

This Fig. 5 shows we are predicting the percentage of ripple effects which may arise due to change in method setVeriable. Various classes like javaObjectBnd, UnMarshall, Utill, XMLFileGeneration will get affected due to this change.

Advantages in the proposed tool called as code change impact analysis tool:

- Independent Impact analysis tool.
- Portable.
- Plug and play system application.
- This tool can be used in tiny as well as massive projects.

CONCLUSION

Change impact analysis tool which helps to determine feasible impacts prior to constructing a modification, that will cut the hazards of receiving on a costly modification since the later the difficulty is exposed the added expensive. This tool offers visibilities into the feasible effects of modifications previous to the changes are applied, in addition it recognize the rate of proposed software modifications. The result report assists developer to build modifications further properly plus show during the effects of modifications. This can be also utilizing to estimate the suitability of considered modifications. Industrial employee can utilize this tool to run "what if" examination on unusual alteration proposals, plus choose the one which help to cut the rate. Program Developers can make a use of this type of tool to indicate the failing of key sections of source code. Software application testers can utilize this tool to determine which program source code sections are impacted in the modifications. This analysis tool can also appropriate in a mixture of computer programming languages for calculating the impact of map modification in innovative program source code.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Wen-Tin Lee, Whan-Yo Deng, Jonathan Lee,Shin-Jie Lee. [2010] Change Impact Analysis with a Goal-Driven Traceability-Based Approach, international journal of intelligent systems, 25:878-908.

[2] Seonah Lee, Sungwon Kang, Sunghun Kim, Matt Staats. [2014] The Impact of View Histories on Edit Recommendations, DOI 10.1109/TSE.2014.2362138, IEEE Transactions on Software Engineering.

[3] Suhaimi Ibrahim, Malcolm Munro. [2009] A requirements traceability to support change impact analysis, Centre For Advanced Software Engineering.

[4] Benjamin Klatt, Martin Kuster, Klaus Krogmann, Oliver Burkhardt. A Change Impact Analysis Case Study: Replacing the Input Data Model of SoMoX, FZI Research Center for Information Technology.

[5] Samuel ajila. [1995] Software Maintenance: An Approach to Impact Analysis of Objects Change, software—practice and experience, 25(10): 1155–1181.

[6] Tor Stålhane, Vikash Katta, Thor Myklebust, Change Impact Analysis in Agile Development, Norwegian University of Science and Technology, OECD Halden Reactor Project and Norwegian University of Science and Technology.

[7] Lile Hattori, Gilson dos Santos Jr, Fernando Cardoso, Marcus Sampaio. Mining Software Repositories for Software Change Impact Analysis: A Case Study, XXIII Simpósio Brasileiro de Banco de Dados.

[8] Malcom Gethers, Bogdan Dit, Huzefa Kagdi, Denys Poshyvanyk. [2009] Impact Analysis for Managing Software Changes, https://sites.google.com/site/asergrp/dmse.

[9] Suhaimi Ibrahim, Norbik Bashah Idris, Malcolm Munro, and Aziz Deraman. [2005] Integrating Software Traceability for Change Impact Analysis, the International Arab Journal of Information Technology, 2(4).

[10] SM Ghosh, HR Sharma, V Mohabay. [2011] Study of Impact Analysis of Software Requirement Change in SAP ERP, International Journal of Advanced Science and Technology 33.

[11] Neha Rungta, Suzette Person, Joshua Branchaud. [2012] A Change Impact Analysis to Characterize Evolving Program Behaviors, 978-1-4673-2312-3/12/$31.00 @IEEE.

[12] Mark Sherriff and Laurie Williams. [2008] Empirical Software Change Impact Analysis using Singular Value Decomposition.

[13] Bixin Li, Xiaobing Sun, Hareton Leung, Sai Zhang. [2012] A survey of code-based change impact analysis techniques, software testing, verification and reliability Softw. Test. Verif. Reliab. Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/stvr.1475.

ISSN: 0976-3104
SPECIAL ISSUE : Computer Science

ARTICLE

A CRYPTOGRAPHICALY IMPOSED DCP-ABE-M SCHEME WITH ATTRIBUTE BASED PROXY RE-ENCRYPTION AND KEYWORD SEARCH IN UNTRUSTED PUBLIC CLOUD

Suraj U Rasal[1], Varsha S Rasal[1], Shraddha T Shelar[2]

[1*]*Computer Engineering Department, Bharati Vidyapeeth University College of Engineering Pune, INDIA*

[1]*Department of Computer Science & Engineering, Nehru College of Engineering & Research Center, Thrissur, Kerala, INDIA*

[2]*Department of Information Technology, DY Patil College of Engineering Akurdi, Pune, INDIA*

ABSTRACT

Encryption on the user data before outsourcing it into the cloud is an inevitable task. Keyword searching is an influential technique which enables the cloud to conduct keyword searching on the encrypted data which allows performing various search operations on re-encrypted data for rapid data retrieval. Most of the existing schemes had focused on single user scenario. The proposed system concentrates on the multiple senders and multiple user scenarios. Here we have merged the concepts of DCP-ABE-M (Decentralized Cipher Policy Attribute Based Encryption with mediator) with ABRKS (Attribute Based proxy Re-encryption with Keyword Search) which enables the option of advance secure re-encryption and keyword searching in unsecure public cloud. The proposed DCP-ABE-M-ABRKS (Decentralized Cipher Policy Attribute Based Encryption with mediator - Attribute Based proxy Re-encryption with Keyword Search) system provides some special features such as: (i). data owner can ask the cloud to conduct keyword search, on his encrypted data by using the user given search token, (ii). Cloud re-encrypts the available cipher text by using a cryptographically enforced DCP-ABE-M technique which contains additional features for data security. Hence, the proposed system is more reliable than the existing systems.

KEY WORDS
Public key Encryption, Proxy Re-encryption, Keyword Search, cloud storage, DCP-ABE, Mediator, OTP.

Received: 16 December 2016
Accepted: 15 January 2016
Published: 15 February 2017

***Corresponding Author**
Email:
surasal@bvucoep.edu.in
Tel.: +918793000079

INTRODUCTION

Eminently, cloud computing is the prominent platform which assemble tremendous computational resources and make them available as a service to varies users. The cloud users can store their data and can enjoy the promising properties of cloud. Predominantly, there are three types of clouds: 1) public cloud: they are owned and operated by companies and provide rapid services to users, 2) private cloud: they are owned and operated by a single organization and provides rapid services for specific authorized user, 3) hybrid cloud: it's a combination of public cloud and private cloud where the users can enjoy the facilities of public cloud also. Public cloud is more reliable, cost saving and elastic in nature than private cloud. Public cloud provides some benefits such as: SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service). Inorder to store the data into the cloud the user must encrypt the data before outsourcing it into the cloud for safeguarding the data privacy.

The most important benefits of using cloud are low maintenance cost, pervasive accessing and storage flexibility etc. But on the other hand, cloud storage faces many troubles against service quality [1][2] and vulnerabilities.

In order to ameliorate the data privacy the existing scheme [4] added public key encryption techniques in cloud storage which allows storing the encoded data in the cloud. That is for safeguarding the data privacy user should encrypt the data before outsourcing it into the cloud and this scheme had given a token to a designated user so that except that user no one will be able to decrypt the data. But in the case of multicasting this scheme fails. In order to solve this problem proxy re-encryption technique is used [3]. In this scheme a semi trusted proxy re-encrypt the cipher text for remaining users. Yanfeng Shi, Jiqiang Liu proposed a scheme Attribute-Based Proxy Re-Encryption with Keyword Search [5] which allows attribute based re-encryption for the user by using keyword search for fast access.

Here we have embedded the DCP-ABE-M [6] public key encryption scheme, key word search technique with proxy re-encryption [5] and applied it on cloud storage for improving security and faster response.

MATERIALS AND METHODS

This chapter shows the summarized features of the most relevant techniques, proxy re-encryption with keyword search, attribute-based encryption, attribute-based encryption with keyword search and attribute-based proxy re-encryption, DCP-ABE, DCP-ABE-M with OTP.

Public key Encryption: ABE

Inorder to avoid the limitations of IBE (Identity Based Encryption) Sahai and waters had proposed a scheme called ABE (Attribute Based Encryption) [7]. In ABE scheme focal power will be responsible for the global initialization of ABE frame work. Focal power or central authority screens the arrangement of all

attributes of various users and then allocate the mystery keys to users based upon their priority. Prominently, the client can decode the encrypted data if and only if there is a match between the tracts which is attached with the cipher text and the user holding qualities. Henceforth, this system can be used as an essential center in the exploration group.

Attribute-based encryption with keyword search

Zheng QJ, Xu SH had proposed the concept of ABK (Attribute Based Keyword search) [8]. Then this work is extended by Sun WH, Yu SC [9] for increasing the data retrieval speed. This scheme allows the data owner to provide some keywords for their encrypted data so that the authorized users can easily access the cipher text by using those keywords.

Attribute-based proxy re-encryption

Boneh et al. had proposed the first PEKS (Public key Encryption Keyword Search) scheme in 2004 [4]. The concept of designated tester (dPEKS) is used in this scheme, which will only allow a designated server to run the test function (dTest). This scheme fails when the sender wants to send the same message to multiple people. Further, this problem is solved in [3] which introduced the concept of proxy re-encryption technique at first time. Proxy re-encryption scheme allows a semi trusted proxy to encrypt a cipher text or already encoded code into another cipher text of same message by using sender's public key and some special information. This concept is further extended in [10][11][12][13]. But the limitation of this scheme is, it requires more accessing time or the users may face more difficulty in accessing of cipher text.

Proxy Re-encryption with keyword search

The drawback of the scheme [14] can be avoided by adding the concept of keyword searching in it. Inorder to solve the problem of previously existing system Shao J, Cao ZF, Liang XH introduced a new scheme Proxy Re-encryption with Keyword Search [15]. This scheme allows the data owner to provide the option of keywords to other users. Then this concept is further extended by Yau WC, Phan RCW in [16][17][18].

Attribute based proxy re-encryption and keyword search

Yanfeng Shi, Jiqiang Liu introduced the concept ABRKS (Attribute Based Proxy Re-encryption with Keyword Search) [5] in which allows the data owner to conduct the keyword search on the re-encrypted data for other users. Here, the privacy of the keyword is also secured. The following [Fig. 1] represents the re-encryption process.

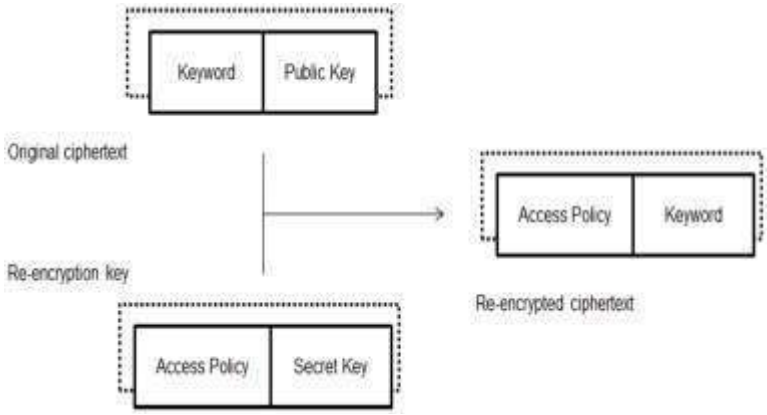


Fig. 1: Enabling the Idea of keyword searching on re-encrypted data

DCP-ABE

Inorder to eliminate the dependency on the focal, Jinguang Han proposed a new scheme called DCP-ABE (Decentralized Cipher text Policy Attribute Based Encryption) scheme [19] which eliminates the drawbacks of existing public key encryption scheme. This system doesn't require a focal for monitoring various components of the system. Multiple independent powers are used in this system which generates secret key without knowing the user attributes, GID and stores the key in a logical manner inorder to secure its privacy. It takes multiple sub secret keys from various powers and they are merged with each other to form the main secret key for a single user. If any power fails then it will be difficult to get the secret key, this is the drawback of DCP-ABE scheme.

DCP-ABE-M

The drawback of the existing scheme can be eliminated through the newly proposed scheme DCP-ABE-M (Decentralized Cipher text Policy Attribute Based Encryption with Mediator) [6]. In this scheme multiple independent powers and mediators are used which are based upon specific attributes. The system allocates a single mediator and power for a single user, who generates secret key for the user based upon the user attributes. The secret key will be split and stored differently in mediator and in power. So that the failure of any component will not affect the entire system.

Table 1: Survey on the existing schemes and its solution

Scheme	Proxy re-encryption	Keyword Search	Access Control	Accessing Speed
PRES [15-18]	✓	✓	x	x
ABE [7]	x	x	✓	x
ABKS [8-9]	x	✓	✓	✓
ABPRE [10-14]	✓	x	✓	x
ABRKS [5]	✓	✓	✓	✓
DCP-ABE [19]	x	x	✓	x
DCP-ABE-M [6]	x	x	✓	x
Proposed system	✓	✓	✓	✓

RESULTS

In the proposed system we have integrated the concept of DCP-ABE-M with proxy re-encryption and keyword searching which enables fast and secure data sharing in cloud. There are 4 factors which plays an important role in this system.

- Focal
- Data owner
- Authorized users
- Cloud

Focal is the central power of the system, which is accountable for the key distribution. When the data owner wishes to outsource his data into the cloud, subsequently the system performs various types of encryption on the data for securing its privacy. In this system data is encrypted two times before storing it in to the cloud. Following [Fig. 2] represents the architecture of the proposed system.

Primary encryption

Before outsourcing of data into the cloud, data owner attach some keywords with data file inorder to get the facility of keyword searching. ABE (Attribute Based Encryption) is used here for primary encryption of data. ABE utilizes the attribute set of user for secret key generation. Here U_A indicates user attribute set, A_n specifies user attribute, S_{K1} represents secret key for first encryption, D_F defines data file, K_I specifies keyword index and E_1 specifies single encrypted file or cipher text.

$U_A = \{A_1, A_2, A_3 \ldots\ldots\ldots A_n\}$
$U_A \rightarrow S_{K1}$
$S_{K1} (D_F + K_I) \rightarrow E_1$
$(S_{K1}.D_F) + (S_{K1}. K_I) \rightarrow E_1$

Here P_E indicates primary encryption,

$P_E \rightarrow D_F + K_I$
$\rightarrow E_1 (D_F + K_I)$
$\rightarrow (E_1.D_F) + (E_1. K_I)$

Subsequently, the encrypted data file is outsourced into the cloud for storage. Cloud is operated by using a semi trusted proxy server. Then the data owner sends a request for keyword search to proxy based upon available token and data. Again data owner sends a request of re-encryption to proxy. Inorder to search

over the re-encrypted data a keyword is given to proxy. By using this keyword, data owner can search on the re-encrypted data for its future retrieval.
Server sends the newly arrived encrypted data file into token recognizer for authentication and verification of token. T_R checks the token and data for verification of user type.

$U_T = E\ K_I$
$T_R \rightarrow \quad U_T \in A_U$
A_U = {Set of authorized user tokens}
If $\quad U_T \in A_U \rightarrow$ authorized user data
Else $U_T \quad A_U \rightarrow$ unauthorized user data

Where, U_T is the user token, $E\ K_I$ defines encrypted keyword index, T_R represents token reorganizer and A_U is the authorized user token set. Here, if newly arrived user token U_T is an element of authorized user token set then that specific user is an authorized or identified user else the user is a fake user. The following flow chart [Fig. 3] represents the processing of this proposed system.

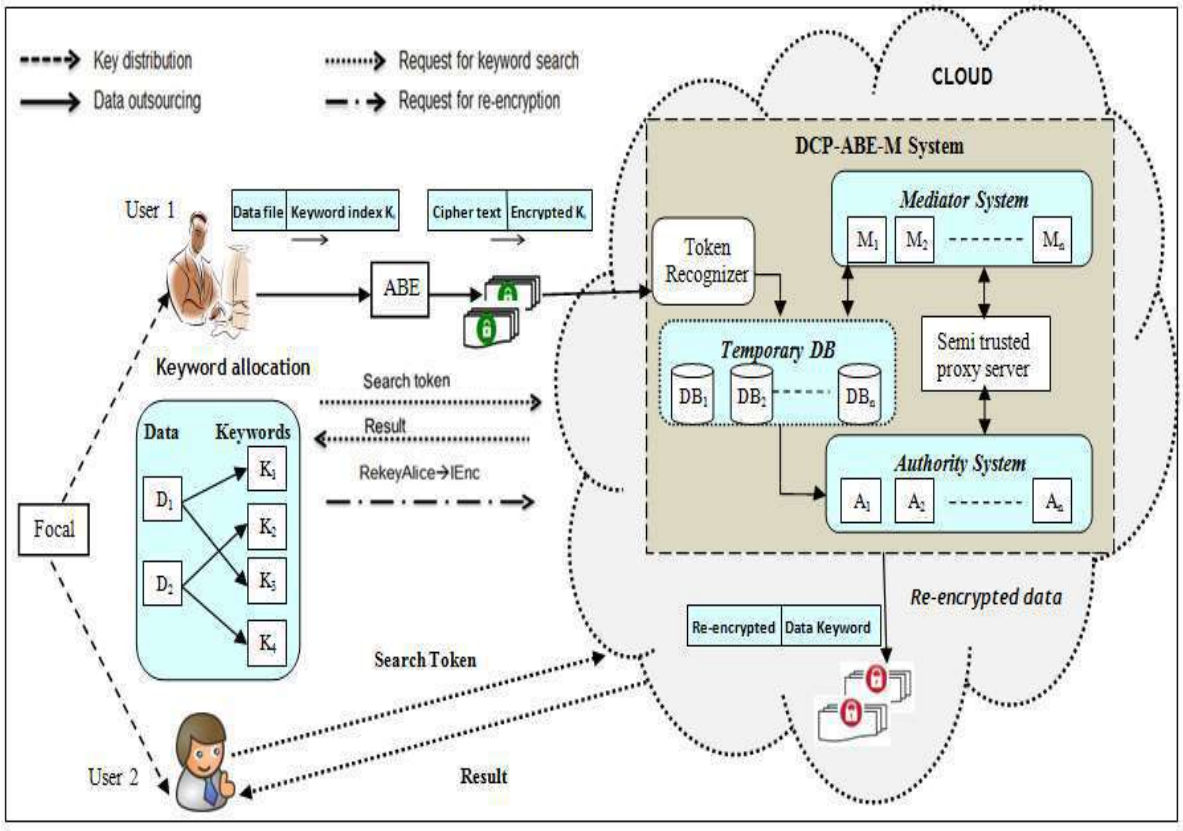


Fig. 2: Architecture of DCP-ABE-M-ABRKS system.

Secondary encryption

The proposed system has used DCP-ABE-M technique for performing secondary encryption. DCP-ABE-M system mainly contains 3 components.

- Mediator system
- Authority system
- Temporary data base

Mediator system contains multiple independent mediators who verifies and checks the attributes of incoming token for the further redirection toward specific authority. Each mediator and authority is based upon some special attributes or characteristics and they takes user data based upon the user attributes or character which matches with their own character. The main purpose of mediator is for storing the half secret key for the security improvement of system.

Here, the token reorganizer verifies the incoming token and stores the data into a specific temporary DB. For each and every user, our system builds a temporary DB and deletes the DB after secret key creation.

Then based upon user type or attribute, our system directs the user towards the mediator which have some matching characteristics of user. For example: business related user is directed towards a mediator who is based upon business attribute. Mediator again verifies all the user details by extracting it from temp DB of that user and direct towards matching authority.

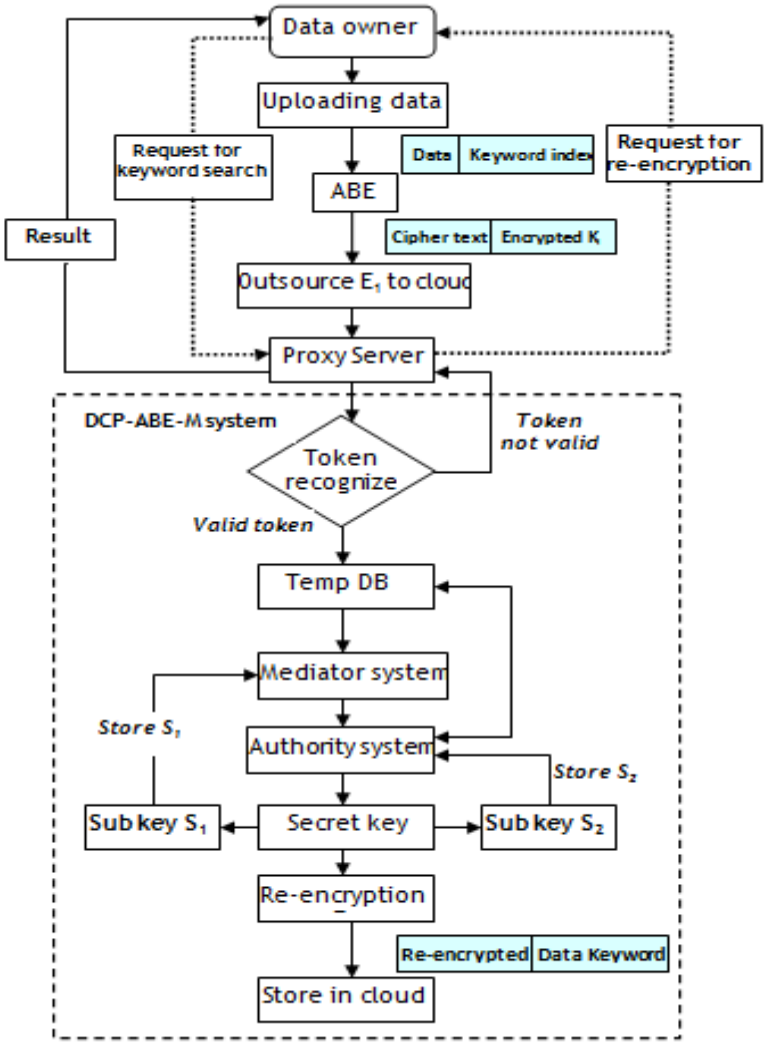


Fig. 3: Flow chart for DCP-ABE-M-ABPKS system processing.

Authority system contains number of independent authorities who generates secret keys for the user. By extracting the details form temp DB, authority generates the secret key for the user which is used for the data encryption.

$S_E \rightarrow [P_E]\ S_{K2}$ ..... (1)
$S_{K2} = U_A + U_T + D_A$

Here, S_E indicates secondary encryption and S_{K2}defines second secret key used for the re encryption. S_{K2} is generated by using user attributes U_A, user token U_T and data attribute D_A.

$U_A = \{U_1, U_2, U_3 \ldots\ldots U_n\}$
$U_T = E_1 K_I$
$D_A = \{D_1, D_2, D_3 \ldots\ldots D_n\}$
$U_A \cup U_T \cup D_A = S_{K2}$
We have
$P_E \rightarrow (E_1. D_F + E_1. K_I)$ ..... (2)
Use (2) in (1) this in S_E
$S_E \rightarrow (E_1. D_F + E_1. K_I)\ S_{K2}$
$S_E \rightarrow (E_1. D_F. S_{K2} + E_1. K_I. S_{K2})$
$S_E \rightarrow E_2$

E_2 specifies re-encrypted data file. By using the secret key S_{K2} , E_1 is re-encrypted. After the encryption data is stored along with keyword by applying some access policy and then subsequently the temp DB is deleted. Prominently, the available secret key S_{K2} is divided into two parts and one part is stored in the mediator and another part is stored in the authority.
When any user tries to access data, focal check the user attributes for authentication and then if the user is an authorized user then the system allows him to access the data and provides specific keyword.

CONCLUSION

The proposed scheme is an integration of technologies in the area of cryptography. The concept of DCP-ABE-M-ABRKS technique can be used in the unsecure public cloud which provides the features like: secure re-encryption, keyword searching facility on encrypted data, securing the privacy of the keyword. Hence our system is more secure and faster in data retrieval than the other existing schemes.

CONFLICT OF INTEREST
There is no any form of conflict of interest

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Ding S, Yang SL, Zhang YT, Liang CY, Xia CY. [2014] Combining qos prediction and customer satisfaction estimation to solve cloud service trustworthiness evaluation problems. Knowl-Based Syst 56: 216–225.

[2] Ding S, Xia CY, Zhou KL, Yang SL, Shang JS. [2014] Decision support for personalized cloud service selection through multi-attribute trustworthiness evaluation. PloS one 9(6): e97762.

[3] Blaze M, Bleumer G., Strauss M. [1998] Divertible protocols and atomic proxy cryptography. *In: Nyberg*, K. (ed.) EUROCRYPT 1998. LNCS, vol. 1403, pp. 127–144. Springer, Heidelberg.

[4] Boneh D, Crescenzo GD, Ostrovsky R, Persiano G. [2004] Public key encryption with keyword search. *In: Cachin C, Camenisch, J.L.* (eds.) EUROCRYPT 2004. LNCS, vol. 3027:. 506–522. Springer, Heidelberg.

[5] Yanfeng Shi, Jiqiang Liu, Zhen Han. [2014] Attribute-Based Proxy Re-Encryption with Keyword Search. PloS one 9(12): e116325.

[6] Varsha Thanaji Mulik, Shinu A, Suraj Rasal. [2016] Privacy Preserving Through Mediator in Decentralized Ciphertext policy Attribute Based Encryption, IJRET: International Journal of Research in Engineering and Technology, 05 (06) | Jun.

[7] Sahai and B. Waters. [2005] Fuzzy identity-based encryption, in *Advances in Cryptology* (Lecture Notes in Computer Science), vol. 3494. Heidelberg, Germany: Springer-Verlag, pp. 457–473.

[8] Zheng QJ, Xu SH, Ateniese G. [2014] VABKS: verifiable attribute-based keyword search over outsourced encrypted data. In: 2014 IEEE Conference on Computer Communications, INFOCOM 2014, Toronto, Canada, April 27 - May 2. pp. 522–530.

[9] Sun WH, Yu SC, Lou WJ, Hou YT, Li H. [2014] Protecting your right: Attribute-based keyword search with fine-grained owner-enforced search authorization in the cloud. In: 2014 IEEE Conference on Computer Communications, INFOCOM 2014, Toronto, Canada, April 27 - May 2, 2014. pp. 226–234.

[10] Li KY, Wang JF, Zhang YH, Ma H . [2014] Key policy attribute-based proxy re-encryption and rcca secure scheme. Journal of Internet Services and Information Security (JISIS) 4: 70–82.

[11] Liang XH, Cao ZF, Lin H, Shao J. [2009] Attribute based proxy re-encryption with delegating Capabilities. In: Proceedings of the 4th International Symposium on Information, Computer, and Communications Security. ACM, pp. 276–286.

[12] Luo S, Hu JB, Chen Z. [2010] Ciphertext policy attribute-based proxy re-encryption. *In: Information and Communications* Security, Springer. pp. 401–415.

[13] Mizuno T, Doi H. [2011] Hybrid proxy re-encryption scheme for attribute-based encryption. *In: Information Security and Cryptology*. Springer, pp. 288–302.

[14] Guo SQ, Zeng YP, Wei J, Xu QL. [2008] Attribute-based re-encryption scheme in the standard model. Wuhan University Journal of Natural Sciences 13: 621–625.

[15] Shao J, Cao ZF, Liang XH, Lin H. [2010] Proxy re-encryption with keyword search. Information Sciences 180: 2576–2587.

[16] Yau WC, Phan RCW, Heng SH, Goi BM. [2010], Proxy re-encryption with keyword search: new definitions and algorithms. *In: Security Technology, Disaster Recovery and Business Continuity,* Springer. pp. 149–160.

[17] Fang LM, Susilo W, Ge CP, Wang JD. [2012] Chosen-ciphertext secure anonymous conditional proxy re-encryption with keyword search. Theoretical Computer Science 462: 39–58.

[18] Wang XA, Huang XY, Yang XY, Liu LF, Wu XG. [2012] Further observation on proxy re-encryption with keyword search. Journal of Systems and Software 85: 643–654.

[19] Jinguang han, willy susilo, yi mu, jianying zhou, Man ho allen au. [2015] improving privacy and security in decentralized ciphertext-policy attribute-based encryption, ieee transactions on *information forensics and security*. 10(3) (1):665-678.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

COMPARATIVE STUDY OF ROUTING PROTOCOLS IN WIRELESS MESH NETWORKS

B Sathyasri[1*], EN Ganesh[2], P Senthil Kumar[3]

[1]*Department of Electronic and Communications Engineering, VEL TECH, Chennai, INDIA*
[2] *Department of Electronic and Communications Engineering, Saveetha Engineering College, Chennai*
[3]*Department of Computer Science and Engineering, S.K.R Engineering College, Chennai, INDIA*

ABSTRACT

Fixed and mobile wireless devices are provided reliable access through wireless mesh networks. Traffic between mesh nodes and internet is a challenging task and is routed over mesh gateways. Path from mesh node to internet node is called forward path and mesh node has to be provided with route information of only one destination (i.e. gateway). Whereas on backward path, internet to mesh node, an individual route for every mesh node is necessary. Therefore in this project, we investigate protocols for backward path routing. The three protocols for backward path routing, AODV- a reactive routing protocol, FBR - a proactive routing protocol and GSR- source routing protocol are compared. Our results indicate that FBR has highest packet delivery ratio but is not scalable to network size. Extended AODV seems to be neither scalable nor does it achieve a high packet delivery ratio. The efficient protocol GSR is most scalable to network size and also achieves a high packet delivery ratio.

KEY WORDS
ACK, Ad- hoc on demand Distance Vector, FBR, Gateway Source Routing, RREP, RREQ, RERR, Wireless Mesh Network

INTRODUCTION

A wireless mesh network (WMN) is a telecommunications system built up of radio nodes standardized in a mesh topology. Wireless mesh networks regularly subsist of mesh clients, mesh routers and portal. The mesh clients are orderly laptops, cell phones and alternate wireless devices although the mesh routers progressive traffic to and from the portal which may, but need not, connect to the internet. The scope area of the radio nodes working as a single network is sometimes called a mesh cloud. Access to this mesh cloud is inferior on the radio nodes working in harmony with each other to conceive a radio network. A mesh network is dependable and offers redundancy. When one node can no longer complete, the rest of the nodes can still interact with each other, directly or through one or more transitional nodes.

A wireless mesh network can be seen as a particular type of wireless ad-hoc network. A wireless mesh network usually has a more prepared composition, and may be expanded to contribute dynamic and cost effective connectedness over a certain geographic area. An ad-hoc network, on the alternate hand, is formed ad-hoc when wireless devices come within intercommunication specifies of each alternate. The mesh routers may be mobile, and be moved according to specific interests arising in the network. Regularly the mesh routers are not specified in terms of resources related to alternate nodes in the network and thus can be overworked to fulfill more resource intensive functions.

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

The characteristics of WMNs are explained as follows

Multi-hop wireless network

A detached to establish WMNs is to line-of-sight (NLOS) connectedness among the end users without direct line-of-sight (LOS) association. To meet these demands, the mesh-style multi-hopping is fundamental, which accomplish higher throughput without endure effective radio range via shorter link distances, less interference between the nodes, and more efficient frequency reiterate.

Hold for ad-hoc networking, and potential of self-forming, self-healing, and self-organization

WMN's strengthen network performance, because of tensile network architecture, easy distribution and configuration, fault tolerance, and mesh connectedness, i.e., multipoint-to-multipoint communications. Due to these features, WMN's have low upfront investment requirement, and the network can grow gradual as needed.

Mobility dependency on the type of mesh nodes: Mesh routers usually have minimal mobility, while mesh clients can be stationary or mobile nodes.

Various types of network approach

In WMNs, to get data from an end user to a node in a major network such as the Internet access to the internet and end-to-end communications are guided. In addition, the synthesis of WMNs with other wireless

***Corresponding Author**
B Sathyasri, Department of Electronic and Communications Engineering, VEL TECH, Chennai, INDIA

networks and afford services to end-users of these networks can be adept through WMNs. Dependence of power-consumption motive on the type of mesh nodes. Mesh routers usually do not have strict motive on power consumption. However, mesh clients may desire power active protocols.

Congeniality and interoperability with current wireless networks

For example, WMNs built based on IEEE 802.11 technologies must be suitable with IEEE 802.11 standards in the impression of supporting both mesh capable and conventional Wi-Fi clients. Such WMNs also need to be inter-operable with other wireless networks such as Wi-MAX, Zig-Bee, and cellular networks. Based on their quality, WMNs are generally investigated as a type of ad-hoc networks due to the lack of wired infrastructure that remain in cellular or Wi-Fi networks through distribution of base stations or access points. While ad-hoc networking techniques are enforced by WMNs, the additional effectiveness require more sophisticated algorithms and design principles for the recognition of WMNs. More specifically, rather of being a type of ad-hoc networking, WMNs aim to diversify the capabilities of ad-hoc networks. Therefore, ad-hoc networks can absolutely be considered as a subset of WMNs. To illustrate this point, the inequality between WMNs and ad-hoc networks are zoned below. In this comparison, the hybrid architecture is considered, since it comprises all the advantages of WMNs.

Wireless framework/resolution

WMNs subsist of a wireless backbone with mesh routers. The wireless backbone provides large scope, connectedness, and robustness in the wireless domain. However, the connectedness in ad-hoc networks depends on the individual subscriptions of end-users which may not be stable

Combination

WMNs hold current clients that use the same radio technologies as a maze router. This is adept through a host-routing function applicable in mesh routers. WMNs also facilitate integration of various existing networks such as Wi-Fi, the internet, the cellular and sensor networks through gateway/bridge functionalities in the mesh routers. Therefore, users in one network are sustaining with services in other networks, over the use of the wireless frame work. The combined wireless networks over WMNs relate the internet backbone, seeing that the physical location of network nodes becomes limited than the capacity and network topology.

Adherence routing and frame work

In ad-hoc networks, ultimate consumer devices also achieve routing and structural functionalities for all other nodes. However, WMNs contain maze routers for these functionalities. Hence, the load on end-user devices is extremely decreased, which provides lower energy utilization and high-end application capabilities to possibly mobile and energy strained end-users. Moreover, the end-user requirements are confined which decreases the cost of devices that can be used in WMNs.

Collective radios

Mesh routers can be implemented with collective radios to perform routing and access functionalities. This provides the segregation of two main types of traffic in the wireless domain. While routing and frame work are expanse between maze routers, the access to the network by end users can be carried out on a different radio. This extremely improves the expanse of the network. On the other hand, in ad-hoc networks, these functionalities are achieved in the same channel, and as a result, the execution decreases.

Portability

A foundational problem of multi-hop wireless networks is the confined scalability and reduction of completion with expanding path lengths, i.e. number of hops. This constrain is mainly due to co-channel interference as well as the certainty that IEEE 802.11 interfaces do not hold full-duplex application, i.e. synchronous transmission and reception of data. One access to run over this problem is to use multi-homed (multi-radio) nodes, with radio transceivers tuned to orthogonal channels. Multi-homed nodes have extremely increased capacity, due to decomposition interference and the ability to perform full-duplex communication, which is not sustained by single radio nodes. In addition to degradation interference via increased channel diversity, these appended interfaces can be used to create multiple synchronous links.

Ad-hoc On Demand Distance Vector Routing (AODV) is a peculiar algorithm for the performance of ad-hoc networks. Each Mobile Host enforce as a specialized router and routes are received as required (i.e. on demand), with little or no reliance on cyclic advertisements. This routing algorithm is quite useful for a dynamic self-starting network as required by users desiring to take advantage of ad-hoc networks. AODV provides circumference free routes even while retrieve broken links. Here, this algorithm scales to enormous populations of mobile nodes desiring to form ad-hoc networks. As compared to DSDV and other algorithms which reserve moderately amend routes to all target in the ad-hoc network, this algorithm has quick

response to link breakage in active routes and also reduces memory requirements and causeless reproduction.

ROUTING PROTOCOL

A routing protocol establishes how routers communicate with each other, propagating information that qualify them to prefer routes between any two nodes on a computer network. Routing algorithms induce the specific choice of route. Each router has a preceding knowledge only of networks attached to it without deviation. A routing protocol claim this information first among extant neighbours, and then every place the network. This way, routers benefits attainments of the topology of the network .Routing protocols were conceive for routers. These protocols have been accomplished to allow the replacement of routing tables, or known networks, surrounded by routers. There are a lot of different routing protocols, each one designed for particular network sizes.

Types of Routing

The router learns about remote networks from neighbour routers or from an controller. The router then constitute a routing table that express how to asset the remote networks. If the network is straightly connected then the router at present knows how to asset to the network. If the networks are not appended, the router must learn how to get to the remote network with one of two static routing (controller not automatically enters the routes in the router's table) or dynamic routing (happens consequently using routing protocols).The routers then restore each other about all the system of connection. If a break occurs e.g. a router decreases, the dynamic routing protocols consequently inform all routers about the break. If static routing is used, then the commander has to restore all changes into all routers and therefore no routing protocol is worn.

Only dynamic routing worn routing protocols, which implement routers to:

- Dynamically determine and control routes.
- Multiply routes.
- Dispose routing to restore other routers.
- Reach accord with other routers about the mesh topology.

Statically programmed routers are not able to determine routes, or circulate routing information to other routers. They circulate data by routes defined by the network commander. A root network is so called because it is a obstacle in the network. There is one route inside of the mesh and another one route is in out of the mesh and, because of this, they can be attained using static routing, hence retaining valuable bandwidth.

Routing protocols is a definite of rules or establish that terminate how routers on a network connect and interchange information with each other, allowing them to select outstanding routes to a outlying network, each router has precedence knowledge only of networks attached to it without deviation. Routers moving routing protocol segment this information first, among instant neighbours, then over the entire network. This way, routers share vision knowledge of the topology of the mesh.

Routing protocols perform several activities, including:

- Network detection.
- Renovate and protect routing tables.

The router which settle at the bottom of a network protect a routing table, which is a ballot of networks and dependent routes known by the router. The routing table carry network addresses for its particular interfaces, which are the straightly connected networks, and also network addresses for remote networks. A distant network is a mesh that can be attained by promoting the packet to a router. Distant networks are combined to the routing table in two ways:

i. By the mesh commander not automatically configure the static routes.
ii. By achieving a dynamic routing protocol.

Dynamic routing protocols are worn by routers to measure information about the reachable of the routers and status of distant networks.

IP ROUTING PROTOCOL

There are various aggressive routing protocols for IP. Here are some of the further common aggressive routing protocols for routing IP packets:

- RIP (Routing Information Protocol)
- IGRP (Interior Gateway Routing Protocol)
- EIGRP (Enhanced Interior Gateway Routing Protocol)
- OSPF (Open Shortest Path First)

- IS-IS (Intermediate System-to-Intermediate System)
- BGP (Border Gateway Protocol)

Advantages of aggressive routing protocol

i. Aggressive routing protocols bring up to date and protect the networks in their routing tables.

ii. Aggressive routing protocols not only arrange a greatest path determination to more networks, they will also induce a new greatest path if the initial path becomes unavailable or there is a change in the topology.

AD-HOC ON-DEMAND LAPSE VECTOR ROUTING PROTOCOL

In November 2001 the MANET (Mobile Ad-hoc Networks) in process cluster for routing of the IEFT community has pronounced the first form of the AODV Routing Protocol (Ad-hoc On Demand Distance Vector). AODV is connected with to the class of Distance Vector Routing Protocols (DV). In a DV every node knows its next node and the line to reach them. A node sustain its own routing table, reserve all nodes in the network, the lapse and the nearest hop to them. If a node is not reachable the lapse to it is set to infinity. Every node sends its near by node regularly its whole routing table. So they can check if there is a helpful route to one more node using this nearest as next hop. When a link split a count-to- infinity could happen.

AODV is an 'on demand routing protocol' with slight delay. That means that routes are only entrenched when needed to diminish traffic on high. AODV holds Unicast, Broadcast and Multicast without any extra protocols. The count-to-infinity and loop problem is determined with continuance numbers and the enrolling of the costs. In AODV every hop has the steady cost of one. The routes age very apace in order to receive the movement of the mobile nodes. Link breakages can locally be replaced very productively. AODV uses IP in a proper way. It pleasures an IP address just as a separate identifier. This can easily be done with mounting the subnet mask to 255.255.255.255. But also accumulate networks are sustained. They are appliance as subnets. Only one router in specific node is important to progress the AODV for the inclusive subnet and provide as a offense gateway. It has to sustain continuance number for the whole subnet and to progressive every package.

In AODV the routing table is explicated by a continuance number to every destination and by time to vital for every passage. It is also explicated by routing flags, the intrusion and a list of vanguard and for antiquated routes the final hop count is reserved.

AODV PROPERTIES

1. AODV detect routes as and when necessary. Does not sustain routes from every one node to every other node.
2. Routes are sustained just as long as specified.
3. Every node sustained its monotonically expanding continuous number, expands every time the node notices change in the next topology.
4. AODV use routing tables to reserve routing information. A Routing table for only single routes .A Routing table for many routes .
5. The route table reserve destination adder, next-hop adder, destination continuous number, life time.
6. For each destination, a node sustained a list of vanguard nodes , to route complete them vanguard nodes help in route preservation.
7. Life-time renovate every time the route is worn .If route not worn within its life time, it elapses.

GATEWAY SOURCE ROUTING PROTOCOL

A source routing protocol re-uses the forward paths that are certified by data packets and reserved on the gateways. These paths are then worn for source routing on the regressive path. With gateway source routing (GSR), the progressive path information from the packets that appear at the gateways is repeated. In the routing header of whole packet, the intermediate hops from the maze node to the gateway are certified. These paths are then reserved in the gateways. To route packets to a maze node, the maze gateway inverts the certified. Progressive path and replica it to the packet header. The gateway then express the packet to the opening node of the recursive path. Each and every node restore the path in the headerby eliminate its entry and progressive the packet to the given next neighbour hop since the packet reaches the terminal. By design, this approach is scalable to the representation of mesh nodes as it establishes no beyond that depends on this number. Only the gateways have to sustained up-to-date routes to single mesh nodes. Also, this path does not upgrade the number of control packets returned between the mesh nodes, and thus decreases the chance of collisions. Surely, GSR depend upon that a packet towards a host in the internet is first sent by a maze node in order to originate the recursive path. HIP and most other addressing mechanisms contain cyclic registration messages from the maze node towards a gateway. Those fluctuating registration messages serve also to induct and sustain the path at the gateway.

In wired networks, the worn has to escort access to wired cables so as to drip transmission. In adverse, the attacker only rights a deserved transceiver to receive wireless signal without being exposed. In wired networks, devices like desktops are constantly static and do not change from one place to another. Hence in wired networks there is no essential to protect users' mobility mode or move pattern, while this delicate information should be kept separate from match in wireless environments. Variously, an match is able to profile users according to their change, and expose or misuse worn based on such information. Finally, providing separate protection for ad-hoc networks with less-power wireless devices and lower-bandwidth network connection is a very difficult task. The tonicity property of the determined path weight is worn to establish a routing protocol that can notify the maximum bandwidth path from each every node to each other destination.

A wireless network may have a lot of dependent routing attacks, in which dropping a malicious conduct of nodes is, current anonymous routing protocols essentially concede anonymity and sectional unlink ability, most of them stroke unbalanced feature of public key crypto systems to attain their goals. Complete unlink ability and un observability are not approved due to short content preservation. Current schemes decline to sustain all content of packets from mugger, so that the mugger can obtain information like packet type and continuous number etc. This information can be worn to express two packets, which split unlink ability and may point to source tread back attacks. Concurrently, powerless packet type and continuous number also make current schemes visible to the adversary. Here, distinct from gateway source routing, an address privacy-maintain routing mechanism is involved that accomplish content observability by exploit anonymous key entrenched based on group signature. The setup of this appliance is simple. Each and every node only has to achieve a cluster signature conform key and an ID-based special key from an offline key server or by a key authority scheme. The gradual routing protocol is then completed in two phases. First, an anonymous key formulation process is performed to compose secret huddle keys. Then an gradual route discovery process is completed to find a route to the destination. This is to maintain all parts of a packet's content, and it is reliant of solutions on transit pattern observability, which thereby add excellent results to the dynamic performance of GSR.

RESULTS AND DISCUSSION

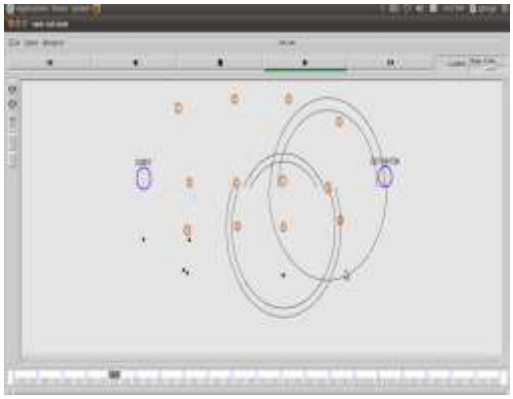


Fig. 1(5): AODV Loss of packets due to link failure or contention

When there is any contention or link failure in the network, there loss of packets occurs thereby reducing the efficiency of the routing protocol. The fig shows the packet loss in the network.

Fig. 1(6): Field Based Routing Protocol Node 8 transmitting packets to the intended destination.

The process of transmission continues till the node having the highest potential is identified, this node is the destination. [Fig. 1(6)] shows the destination receiving packets from node 8.

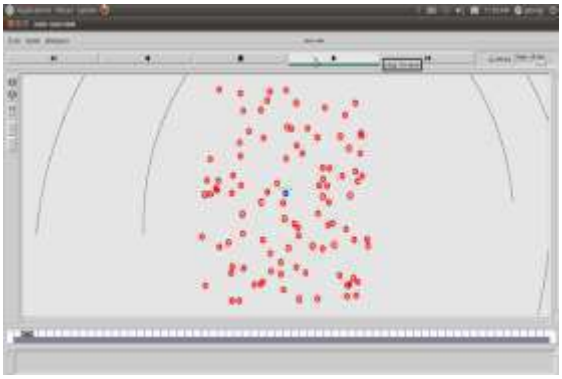


Fig. 1(7): Gateway transmitting packets to the intended destination (node).

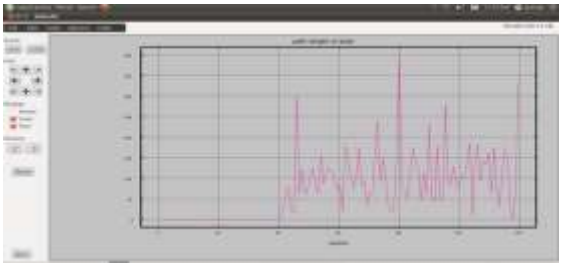


Fig. 1(8): Path length of AODV vs. no. of Packets

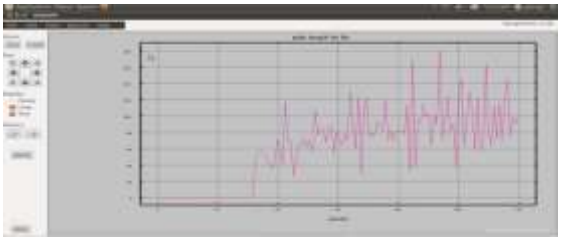


Fig. 1(9): Path length of FBR vs. No. of Packets

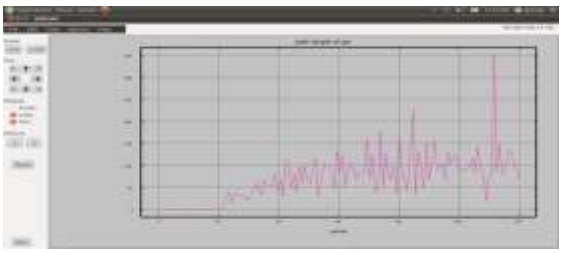


Fig. 1(10): Path length of GSR vs. No. of Packets

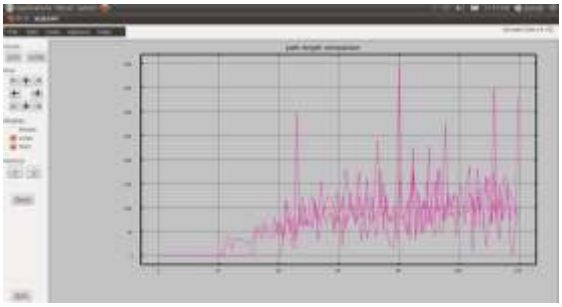


Fig. 1(11): Path length comparison

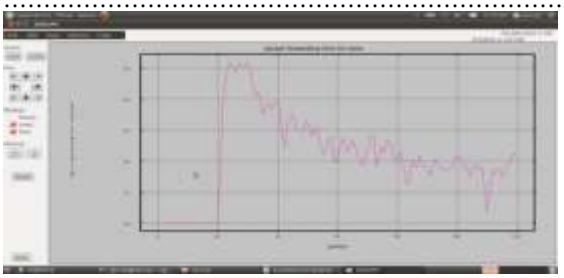


Fig. 1(12): Packet forwarding time of AODV vs. no. of packets

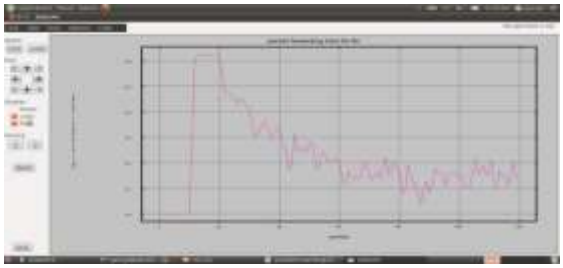


Fig. 1(13): Packet forwarding time of FBR vs. no. of packets

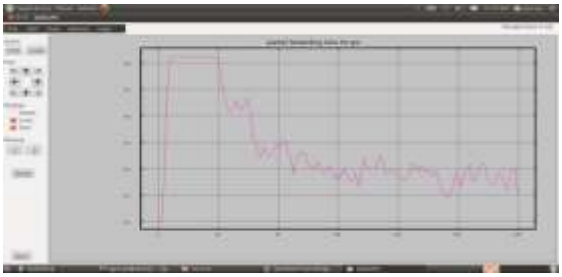


Fig. 1(14): Packet forwarding time of GSR vs. no. of packets

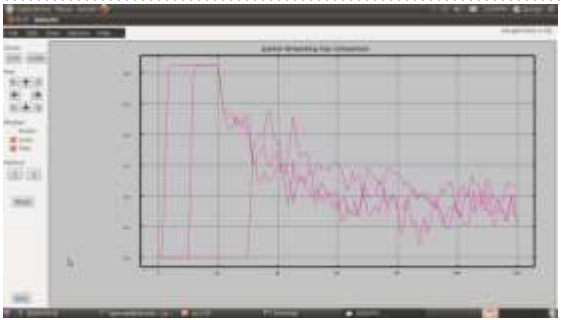
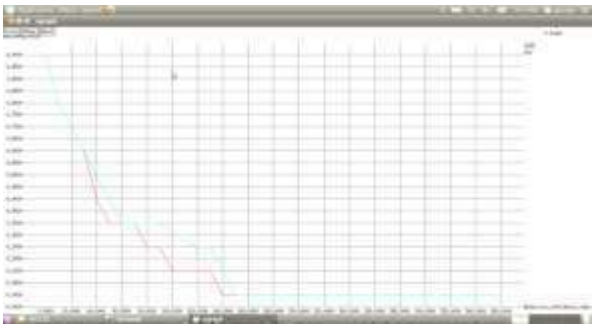


Fig. 1(15): Packet forwarding time comparison

Table 1(1): Representation of protocols



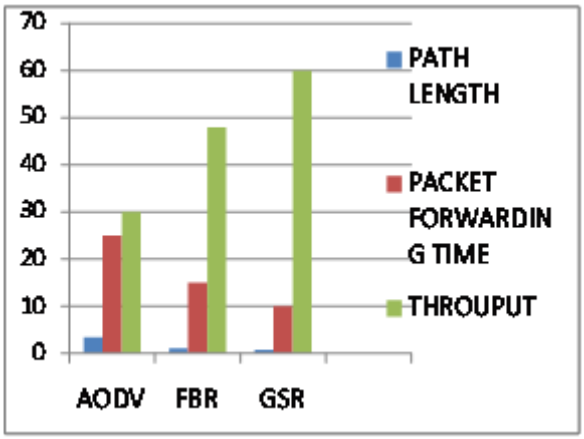


Fig. 1(16): Packet delivery ratio comparison

PROTOCOLS	PATH LENGTH	PACKET FORWARDING TIME	THROUPUT
AODV	3.4	25	30
FBR	1	15	48
GSR	0.75	10	60

Fig 1.17: Comparison for protocol Parameter

CONCLUSION

The main aim of work is increase the overall efficiency of the network. For this purpose we have implemented a protocol called gateway source routing in wireless mesh networks and simulated in NS2 (Network Simulator 2). In this simulation we have created different mesh networks and have transmitted packets using different protocols. The protocols used are, Ad-hoc On demand Distance Vector routing (AODV), Field Based Routing (FBR), Gateway Source Routing (GSR). Hence we analyzed a base paper and found out the best method with which we can improve the efficiency of the network called Gateway Source Routing and have executed the same along with the comparison results of the above mentioned two protocols used for the transmission of the same. This concludes that the gateway acting as the source transmits data efficiently to the intended node using the Gateway Source Routing protocol. The same protocol AODV, FBR and GSR can be implemented in mobile pattern of nodes and the parameters: packet forwarding time, path length and throughput can be simulated using NS2.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] C Perkins, E Belding-Royer and S Das [2003] Ad hoc On-Demand Distance Vector (AODV) Routing, RFC 3561 (Experimental).
[2] T Clausen and P Jacquet [2003] Optimized Link State Routing Protocol.
[3] Vincent Lenders and Martin May and Bernhard Plattner [2006], Density-based vs. Proximity-based Anycast Routing in proceedings of the IEEE Infocom, Barcelona, Spain.
[4] Multi-Linked AODV Routing Protocol for Wireless Mesh Networks Asad Amir Pirzada and Ryan Wishart,[?] Queensland Research Laboratory, Marius Portmann, School of Information Technology and Electrical Engineering, The University of Queensland, Australia.
[5] Vincent Lenders and Martin May and Bernhard Plattner, [2005] "Service Discovery in Mobile Ad Hoc Networks: A Field Theoretic Approach," in Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Taormina, Italy.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

MUSICAL INSTRUMENT SOUND CLASSIFICATION

R Rayar[1], M Anto Bennet[2*], A Nazreen Banu[3], A Sushanthi[4], M Rajasekar[5]

[2]Department of ECE , VEL TECH,Chennai-600062
[1,3,4,5]Department of CSE, Veltech Multitech RR.SR Engineering College, Chennai, INDIA

ABSTRACT

Music instrument classification is essential in music indexing systems. Today Digital Audio Applications is a part of everyday life. Audio in the form of CD's, DVD's and broadcast data, is available in the internet for public access. In this project an automatic music instrument classification system is developed using Discrete Wavelet Transform DWT features. Proximal Support Vector Machine (PSVM) are based on the principle of structural risk minimization. DWT features are extracted from different classes of musical instruments namely flute, guitar, violin and piano. PSVM is trained and tested by using DWT features and the system shows satisfactory results with an accuracy of 89.00%.Index Terms—Discrete Wave Transform, Musical Instrument Sound Classification, Proximal Support vector machine(SVM)

KEY WORDS
Schistosoma nasale, cross H.F cow, Anthiomaline

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

INTRODUCTION

A musical instrument is an instrument created or adapted to is make musical sounds. In general, any object that produces sound can be a musical instrument. It is through purpose that the object becomes a musical instrument. There was history of musical instruments dates to the beginning of human culture. Early musical instruments were used for ritual, such as a trumpet to signal success on the hunt, or a drum in a religious ceremony. Cultures eventually developed composition and performance of melodies for entertainment. Musical instruments evolved in step with changing applications.

Music signals represent a large class of audio data where several sound sources are usually present at the same time. Depending on the genre, the instrument may consist of electric guitars, bass, drums, and vocals or saxophone, piano, strings and percussion, For example, there is a wide variety of instruments in Western music alone, representing different sound production mechanisms and timbre [1],[2]. Automatic recognition of the instruments in recorded music has several direct applications, including music retrieval based on the instrumentation and audio management in recording studios. Even more importantly, sound source recognition and modeling is an essential part of making sense of complex audio signals. When listening to polyphonic music, human listeners are able to perceptually organize the component sounds to their sources, largely based on timbre information. Similarly, source models are an integral part of music transcription and sound separation systems, where the source identity enables the use of source specific models and assumptions and allows the organization of sounds events to "streams" that can be attributed to certain instruments [2] [3].

In addition to practical applications, a system that can automatically classify recordings by genre has significant theoretical musicological interest as well. There is currently a relatively limited understanding of how humans construct musical genres, the mechanisms that they use to classify music and the characteristics that are used to perceive the differences between different genres. A system that could automatically classify music and reveal what musical dimensions it is using to do so would therefore be of great interest.

Low-level signal processing based features are of little use in this respect, something that further emphasizes the importance of studying the use of high-level features This kind of research also has applications beyond the scope of genre classification. The techniques developed for a genre classification system could be adapted for other types of classifications, such as by compositional style or historical period. Once a classification system is implemented, one needs to only modify the particular training recordings and taxonomy that are used in order to perform arbitrary types of classification[9].

One of the most crucial aspects of instrument classification is to find the right feature extraction scheme. During the last few decades, research on audio signal processing has focused on speech recognition, but few features can be directly applied to solve the instrument-classification problem [9],[10]. The identification of the instruments that compose a musical signal has received increasing attention in the last years. Such an interest is fed by the potential benefits that an accurate instrument classifier can bring to other digital audio applications. In particular, musical genre classification can be greatly improved if the instruments present in a given song are known, since this information can be used to narrow down the set of potential musical genres. Sound source separation algorithms can also explore such information, particularly if they deal with underdetermined signals. In this case, the knowledge about the instruments can be used to create instrument specific rules to improve the quality of the sound source separation. Early work in the area was mainly devoted to the identification of instruments in monophonic signals. This

***Corresponding Author**
Email:
bennetmab@gmail.com

problem is in general, less challenging than the polyphonic case, since the instrument to be classified is isolated from the interference of any other sound source. Most of the proposals those deal with general instruments while a few others deal with specific cases like classification of woodwinds and discrimination between piano and guitar [4].

By automatic musical genre classification we mean the classification of music signals into a single unique class based computational analysis of music feature representations. Automatic music genre classification is a fundamental component of music information retrieval systems. The process of genre categorization is described in two steps namely: feature extraction and multiclass classification. In the feature extraction step, extract from the music signals information representing the music. The features extract should be comprehensive (representing the music very well), compact (requiring a small amount of storage), and effective (not requiring much computation for extraction). To meet the first requirement the design has to be made so that the both low-level and high-level information of the music is included [7].

FEATURE EXTRACTION

The Discrete Wavelet Transform

The Wavelet Transform (WT) is a technique for analyzing signals. It was developed as an alternative to the Short Time Fourier Transform (STFT) to overcome problems related to its frequency and time resolution properties. More specifically, unlike the STFT that provides uniform time resolution for all frequencies the DWT provides high time resolution and low frequency resolution for high frequencies and high frequency resolution and low time resolution for low frequencies. In that aspect it is similar to the human ear which exhibits similar time-frequency resolution characteristics.

The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently.

The DWT is defined by the following equation:

$$w(j,k) = \sum_{j}\sum_{k} x(k)2^{-\frac{j}{2}}\Psi(2^{-j}n-k)$$

Where t ψ is a time function with finite energy and fast decay called the mother wavelet. The DWT analysis can be performed using a fast, pyramidal algorithm related to multi rate filter banks. As a multi rate filter bank the DWT can be viewed as a constant Q filter bank with octave spacing between the centers of the filters. Each sub band contains half the samples of the neighboring higher frequency sub band. In the pyramidal algorithm the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive high pass and low pass filtering of the time domain signal and is defined by the following equations.

$$y_{high}[k] = \sum_n x[n]g[2k-n]$$

$$y_{low}[k] = \sum_n x[n]h[2k-n]$$

where $y_{high}[k]$ is the high pass filter
$y_{low}[k]$ is the low pass filter

The output respectively after sub sampling. Because of the down sampling the number of resulting wavelet coefficients is exactly the same as the number of input points. A variety of different wavelet families have been proposed in the literature. In our implementation, the 4 coefficient wavelet family (DAUB4) proposed by Daubechies is used.

Wavelet representation for audio signals

An adaptive DWT and DWPT signal representation is considered in this work because of its highly flexible family of signal representations that may be matched to a given signal and it is well applicable to the task of audio data compression. In this case the audio signal will be divided into overlapping frames of length 2048 samples. [3] When designing the wavelet decomposition considered some restrictions to have compact support wavelets, to create orthogonal translates and dilates of the wavelet (the same number of coefficients than the scaling functions), and to ensure regularity (fast decay of coefficients controlled by

choosing wavelets with large number of vanishing moments). The DWT will act as anorthornormal linear transform. The wavelet transform coefficients are computed recursively using an efficient pyramid algorithm. In particular, the filters given by the decomposition are arranged in a tree structure, where the leaf nodes in this tree correspond to sub bands of the wavelet decomposition. This allows several choices for a basis. This filter bank interpretation of the DWT is useful to take advantage of the large number of vanishing moments. [3]

Wavelets with large number of vanishing moments are useful for this audio compression method, because if a wavelet with a large number of vanishing moments is used, a precise specification of the pass bands of each sub band in the wavelet decomposition is possible. Thus, it can be approximate the critical band division given by the auditory system with this structure and quantization noise power could be integrated over these bands.

Wavelet packet representation

Given a wavelet packet structure, a complete tree structured filter bank is considered. Once I find the "best basis" for this application, a fast implementation exists for determining the coefficients with respect to the basis. However, in the "best basis" approach, they do not subdivide every sub band until the last level. The decision of whether to subdivide is made based on a reasonable criterion according to the application (further decomposition implies less temporal resolution). The cost function, which determines the basis selection algorithm, will be a constrained minimization problem. The idea is to minimize the cost due to the bit rate given the filter bank structure, using as a variable the estimated computational complexity at a particular step of the algorithm, limited by the maximum computations permitted. At every stage, a decision is made whether to decompose the sub band further based on this cost function. Another factor that influences this decomposition is the tradeoff in resolution. If it is decomposed further down, it will sacrifice temporal resolution for frequency resolution.

The last level of decomposition has minimum temporal resolution and has the best frequency resolution. The decision on whether to decompose is carried out top-down instead of bottom-up. Following that way, it is possible to evaluate the signal at a better temporal resolution before the decision to decompose. It is proved in this paper that the proposed algorithm yields the "best basis" (minimum cost) for the given computational complexity and range of temporal resolution.

Feature Extraction & Classification

The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. In order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients are used. That way the statistical characteristics of the "texture" or the "Environmental sound" of the piece can be represented. The distribution of energy in time and frequency for music is different for every environment. The mean of the absolute value of the coefficients in each sub band. These features provide information about the frequency distribution of the audio signal. The standard deviation of the coefficients in each sub band. These features provide information about the amount of change of the frequency distribution. Ratios of the mean values between adjacent sub bands. These features also provide information about the frequency distribution. Points on the wrong side of and as training errors. However, in proximal SVM, all the points not located on the two planes are treated as training errors. In this case the value of training error ξi in [2] may be positive or negative. The second part of the objective function in [2] uses a squared loss function instead of to capture this new notion of error

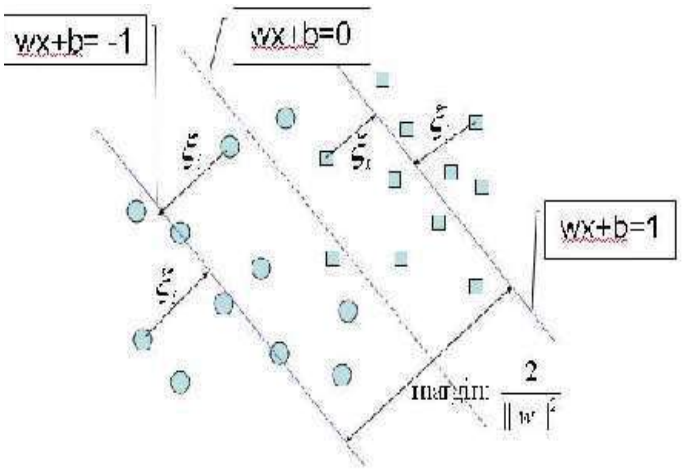


Fig. 1: Proximal SVM

TECHNIQUES

Proximal Support Vector Machine (PSVM)

The proximal SVM also uses a hyper plane as the separating surface between positive and negative training examples. But the parameter w and b are determined by solving the following problem.

$$Min \frac{1}{2}\left(\|w\|^2 + b^2\right) + c\sum_i \xi_i^2 \ldots\ldots\ldots\ldots\ldots 4$$

$$s.t.\ \forall_i y_i(w.x_i + b) + \xi_i = \ldots 5$$

The main difference between standard SVM [1] and proximal SVM [2] is the constraints. Standard SVM employs an inequality constraint where as proximal SVM employs an equality constraint. The intuition of Proximal SVM is shown in Figure 2. We can see that standard SVM only considers . We show the reason why the original proximal SVM is not suitable for classifying unbalanced data in this section. To the unbalanced data, without lose of generality, suppose the amount of positive data is much fewer than the negative data. In this case the total accumulative errors of negative data are much higher than that of positive data. Consequently, the bounding plane will shift towards the direction opposite to the negative data to produce a larger margin at the price of increasing the positive errors. Since the positive data are rare, this action will lower the value of objective function [2]. Then the separating plane will be biased to the positive data and result in a higher precision and a lower recall for the positive training data

The linear multicategory proximal support vector machine (MPSVM)

To motivate our MPSVM we begin with a brief description of the 2-category proximal support machine formulation (Fung & Mangasarian, 2001). We consider the problem, depicted in figure 1, of classifying m points in the n-dimensional real space , represented by the m × n matrix A, according to membership of each point Ai in the class A+ or A− as specified by a given m × m diagonal matrix D with plus ones or minus ones along its diagonal. For this problem, the proximal support vector machine (Fung & Mangasarian,2001) with a linear kernel is given by the following quadratic program with parameter v > 0 and linear equality constraint:

$$\frac{v}{2}\|y\|^2 \text{ ii } \left\|\begin{bmatrix} w \\ \gamma \end{bmatrix}\right\|^2 \ldots\ldots\ldots 6$$

RESULTS

DATASET

The database for the experiments contains 400 samples which are taken from television broadcast database. The recordings are categorized into general classes according to common characteristics of the scenes (100 flute, 100 guitar, 100 violin, 100 piano) and events The categorization of the scenes was somewhat ambiguous, some of the recordings are associated with more than one higher-level class. The recordings are manually labelled and are separated into 1-second, 2-second and 3-second fragments. Every sound signal was stored with some properties that are also the initial conditions and criteria for the well-functioning of the algorithm. The sample database is split into training sets and test sets. In this work on randomly select 80% sounds of each class for the training set. The remaining 20% sounds form the test set. It is have taken different proportion of samples based on class dependency in each category as shown in table.

Category of Musical Instruments sound	No. Of Samples
Flute	96%
Guitar	95%
Violin	94%
Piano	92%

PREPROCEESING

The database is collected from the Television broadcast database. A window size of 16000 samples at 16KHz sampling rate with hop size of 1 second which is used as input for the feature extraction. The training data's are segmented into fixed length overlapped frame (in our experiment 20 ms frames with 10 ms overlapping is used).Since a 16KHz sampling rate is deployed, 20 ms frames consists of 320 values which are converted into 6 dimension for one frame. Here 400 clips used for training data, 40 clips for testing data and each clips must be mono channel.

Fig. 5.2.3: Result displayed

ACOUSTIC DATABASE DESCRIPTOR

CONCLUSION

In this work, "Musical instrument classification" using PSVM modeling techniques and DWT features are extracted to model the music instrument. Features for music instrument are extracted and those models were trained successfully. Music from four different instruments were modeled Using PSVM. in this work,400 database were chosen from television broad caste data, which is considered for training and testing 300 music samples are trained and testing for 100 samples data.

The characteristic of the sound signal collected from the television broad caste database were analyzed. PSVM shows an accuracy of 85% for Flute, 90% for guitar, 88% for violin, 91% for piano ,The results shows the overall performance of accuracy 88.5% using multi class PSVM.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] C Joder,S Essid and G Richard, [2009] Temporal integration for audioclassification with application to musical instrument classification, IEEE Trans. Audio, Speech, Lang. Process. 17(01): 174–186.

[2] T Kitahara, M Goto, K Komatani, T Ogata, and H G Okuno, [2007] Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps, EURASIP J. Appl. Signal Process. 1–15.

[3] P Herrera-Boyer, A Klapuri and M Davy, [2006] Automatic classification of pitched musical instrument sounds, in Signal Processing Methods for Music Transcription, A.KlapuriandM.Davy,Eds. NewYork NY, USA: Springer. 163–200

[4] S Essid, G Richard, and B David, [2006] Instrument recognition in poly-phonic music based on automatic taxonomies, IEEE Trans. Audio,Speech, Lang. Process. 14 (01): 68–80.

[5] A A Livshin and X Rodet, [2004] Musical instrument identification in continuous recordings , in Proc. Int. Conf. Digital Audio Effects, Naples, Italy.
[6] T Kitahara, M Goto, K Komatani, T Ogata and H G Okuno, [2006] Musical instrument recognizer “instrogram” and its application to music retrieval based on instrumentation similarity, in Proc. IEEE Int. Symp. Multimedia. 265–274.
[7] A Eronen, [2001] Comparison of features for musical instrument recognition, in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. 19–22.
[8] G Agostini, M Longari, and E Poolastri, [2003] Musical instrument timbres classification with spectral features, EURASIP J. Appl. Signal Process. 1: 5–14.
[9] A Eronen and A Klapuri, [2000] Musical instrument recognition usingcepstral coefficients and temporal features, in Proc. IEEE Int. Conf Acoust Speech, Signal Process. 753–756.
[10] S. Essid, G. Richard, and B. David, [2006] “Musical instrument recognition bypairwise classification strategies,” IEEE Trans. Audio, Speech, Lang.Process., vol. 14, no. 4, pp. 1401–1412.
[11] S. Essid, G. Richard, and B. David, [2006] “Musical instrument recognitionby pairwise classification strategies,” IEE Trans. Audio, Speech, Lang.Process., vol. 14, no. 4, pp. 1401–1412.
[12] KOSTEK B., CZYZEWSKI A., [2001] "Automatic Recognition of Musical Instrument Sound Further Developments", 110th Audio Eng. Soc. Conv., Amsterdam, 12-15.
[13] T. Kitahara, M. Goto, and H.G. Okuno,[2004] “Category-level identification of non-registered musical instrument sounds,” in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada.
[14] Z. Zhang and B. Schuller, [2012] “Semi-supervised learning helps in sound event classification," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 333-336, IEEE.
[15] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fis-sore, P. Laface, A. Mertins, C. Ris, R. Rose, C. Tyagi, and C. Wellekens, [2007]“Automatic speech recognition and speech variability: A review," Speech Communi-cation, vol. 49, no. 10-11, pp. 763,786.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

IMAGE AND VLSI TECHNIQUES DEVELOPMENT OF EARLY DETECTION OF BREAST CANCER AND DIAGNOSE USING MAMMOGRAMS

B Thamilvalluvan[1], M Anto Bennet[2*], Priyanka Paree Alphonse[3], K Sujithra[4], Soda Chandrasekhar[5], C Kannan[6], D R Thendralarasi[7]

[1]Professor,[2]Asst.professor,[3,4,5,6,7]UG Students,Department of ECE,VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

In this paper we proposed a method to detect the breast cancer accurately in the early stage using image processing and VLSI techniques. In the existing methods cancer affected regions of the mammogram were detected but could not calculate the affected number of cells in that region. In our proposed method a new algorithm has been developed which is able to find the cancer affected regions in the mammogram very accurately and also find the number of cells in the affected region. Initially the mammogram is taken as an input and was read on MATLAB. This mammogram is then converted into a text file based on the gray levels of the mammogram. Then the text file is taken as input in the Xilinx. Using verilog code text file is read and the file is assigned to a variable in binary format. The binary values of the text is read by using the verilog code and the binary values of the text is then compared with the default binary value that is been taken as a binary value of the cancer cell in the verilog code. Then by converting the binary values of the text into decimal value, the cancer cells are detected. Then output of the Xilinx is obtained in the form of waveforms and a text file is created by the output waveforms. Then the decimal values that match with the default decimal values are detected as cancer cells. Finally this text is further converted into image using MATLAB and the number of cancer cells calculated so as to find the affected area.

KEY WORDS
Transductive Support Vector Machine (TSVM), Many Textural Elements (TEXEL), Local Binary Pattern (LBP).

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email
bennetmab@gmail.com

INTRODUCTION

Digital image processing is the use of computer algorithms to perform image processing on digital images. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the build-up of noise and signal distortion during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modelled in the form of multidimensional systems.The identification of objects in an image and this process would probably start with image processing techniques such as noise removal, followed by (low-level) feature extraction to locate lines, regions and possibly areas with certain textures.The clever bit is to interpret collections of these shapes as single objects, e.g. cars on a road, boxes on a conveyor belt or cancerous cells on a microscope slide. One reason this is an AI problem is that an object can appear very different when viewed from different angles or under different lighting. Another problem is deciding what features belong to what object and which are background or shadows etc. The human visual system performs these tasks mostly unconsciously but a computer requires skilful programming and lots of processing power to approach human performance. Manipulation of data in the form of an image through several possible techniques. An image is usually interpreted as a two-dimensional array of brightness values, and is most familiarly represented by such patterns as those of a photographic print, slide, television screen, or movie screen[1-4]. An image can be processed optically or digitally with a computer.To digitally process an image, it is first necessary to reduce the image to a series of numbers that can be manipulated by the computer. Each number representing the brightness value of the image at a particular location is called a picture element, or pixel. A typical digitized image may have 512 × 512 or roughly 250,000 pixels, although much larger images are becoming common. Once the image has been digitized, there are three basic operations that can be performed on it in the computer. For a point operation, a pixel value in the output image depends on a single pixel value in the input image. For local operations, several neighboring pixels in the input image determine the value of an output image pixel. In a global operation, all of the input image pixels contribute to an output image pixel value. One of the techniques in image processing is image segmentation. The inevitable need for image segmentation is to analyze the images in precise manner. It is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, 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. Segmentation is a collection of methods allowing interpreting spatially close parts of the image as objects. Regions (i.e., compact sets) represent spatial closeness naturally and thus are important building steps towards segmentation. The object is everything what is of interest in the image (from the particular Application point of view). The rest of the image is background. The approach is similar to that used in pattern recognition, i.e., division of the image into set of equivalence classes. The purpose of image segmentation is to partition an image into meaningful regions with respect to a particular application. The segmentation is based on measurements taken from the image and might be grey level, colour, texture, depth or motion [5, 6, 7].

MATERIALS AND METHODS

In this paper we proposed a method to detect the breast cancer accurately using image processing and VLSI techniques. Initially the mammogram is taken as an input and was read on MATLAB. This mammogram is then converted into a text file based on the gray levels of the mammogram. Then the text file is taken as input in the Xilinx. Using verilog code text file is read and the file is assigned to a variable in binary format. The binary values of the text is read by using the verilog code and the binary values of the text is then compared with the default binary value that is been taken as a binary value of the cancer cell in the verilog code. Then by converting the binary values of the text into decimal value, the cancer cells are detected. Then output of the Xilinx is obtained in the form of waveforms and a text file is created by the output waveforms. Then the decimal values that match with the default decimal values are detected as cancer cells. Finally this text is further converted into image using MATLAB and the number of cancer cells is calculated so as to find the affected area.

SIGNS OF CANCER IN A MAMMOGRAM

There are several signs of cancer to look for in a mammogram. The primary signs are local distortion of glandular tissue and existents of malignant micro calcifications. These signs may appear alone or together. Skin thickening, skin distortion, retraction of the nipple is considered as secondary signs. The attenuation of tumor may vary from depending upon type of tumor. Benign tumors are usually rounded and have distinct border. While malignant tumors tend to be more irregularly in shape, often speculated and have diffused borders. Normally two breast of one woman are very alike , and cancer of both breast is relatively rare .These facts help radiologist in the search of cancer in a mammogram. Micro calcification is microscopic grains of calcium produced be the cells as the result of some benign or malignant process. Most calcification is the result of some benign process. The may for instance be the rest of broken down cells, a cyst or milk. Benign tumors & malignant calcification differs in shape density and distribution However, it may be found that the predictions possible from the available information are limited because patients with very similar values on all the factors have greatly varying times to relapse. While this problem can be identified by analysis it cannot be addressed by it—the reason for it may be that the risk of relapse depends on some unmeasured factor or is simply subject to great random variation. From a statistical viewpoint, an important feature of the data is the presence of censoring, so that some of the times to relapse are not known exactly, only that they are greater than a certain value. This occurs either because follow-up is no longer available on the patient for some reason (maybe they have died of an unrelated cause), or simply because no relapse has yet occurred, when the survival time is known only to be as long as the total time elapsed since surgery. We assume that censoring times are essentially non-informative, so that for example censoring does not routinely occur just before a relapse is shown in [Fig. 1].

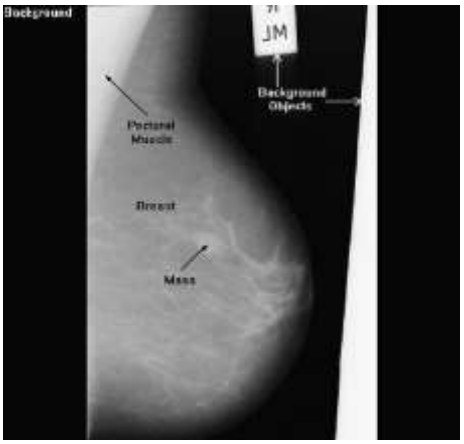


Fig. 1: Illustration of components of mammogram

Texture analysis is widely used for computer vision and image processing for classification and segmentation of image based on a local spatial variation of intensity or color. Existing texture analysis can be classified into statistical and structural methods. Statistical approach computes different properties are suitable if texture primitive sizes are comparable with the pixel sizes. In structural texture analysis method, the texture region is defined to have a constant texture if a set of local statistics or other local properties of the image are constant, slowly varying or approximately periodic. Majority of the existing texture analysis assumes that the texture images are acquired from the same viewpoint (same scale and same orientation). This gives a limitation of these methods. In many practical applications, it is very difficult or impossible to

ensure that images captured have the same translations, rotations or scaling between each other. Texture analysis should be ideally invariant to viewpoint. But recently model based method is also employed which includes autoregressive model, Gaussian Markov random fields, Gibbs random fields, World medal, wavelet model, multichannel Gabor model and steerable pyramid, etc. These models provide more powerful tools for invariant texture analysis. In statistical methods, texture is described by a collection of statistics of selected features. The statistics are broadly classified into first-order statistics, second-order statistics, and higher-order statistics. In structural methods, texture is viewed as consisting of many textural elements (called Texel) arranged according to some placement rules. In model based a texture image is modelled as a probability model or as a linear combination of a set of basic functions. The coefficients of these models are used to characterize texture images.

GREY LEVEL RUN LENGTH MATRIC

Gray level run length matrix is a statistical texture descriptor. This method involves the counting the number of pixels that have the same intensity in a particular direction. Run length is the number of adjacent pixels that have same gray level intensity in a particular direction shown in [Fig. 2]. The texture feature extracted from run length matrices produces great classification results. The texture features that can be extracted by GLRLM includes Short Run Emphasis (SRE), Long Run Emphasis (LRE), Low gray level Run Emphasis (LGLRE), High Gray Level Run Emphasis (HGLRE), Short Run Low Gray Level Emphasis (SRLGLE), Short Run High Gray level Emphasis (SRHGLE), Long Run Low Gray Level Emphasis (LRLGLE), Long Run High gray Level Emphasis (LRHGLE), Gray Level No uniformity (GLN), Run length no uniformity(RLN), Run Percentage (RPERC).

LOCAL BINARY PATTERN OPERATOR

LBP operator combines the characteristics of statistical and structural texture analysis. The LBP operator is used to perform gray scale invariant two-dimensional texture analysis. The LPB operator labels the pixel of an image by thresholding the neighbourhood (i.e. 3 × 3) of each pixel with the centre value and considering the result of this thresholding as a binary number. When all the pixels have been labelled with the corresponding LBP codes, histogram of the labels are computed and used as a texture descriptor. Given a pixel in the image LBP code can be computed by comparing it with its neighbours K is the maximal LBP pattern value. The most important properties of LBP features are computational simplicity and tolerance against the monotonic illumination changes. The basic LBP operator cannot be used for the dominant features of large scale structures. So it is extended to facilitate the analysis of textures with multiple scales by combining neighbourhoods with different sizes.

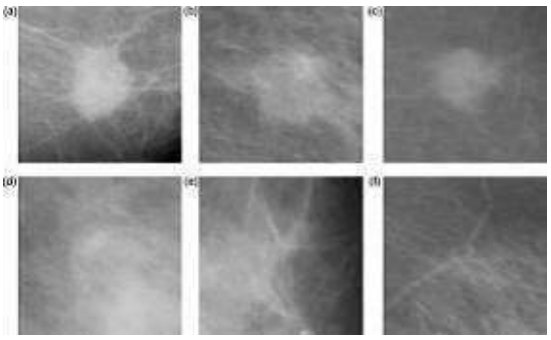


Fig. 2: (a-c) ROI with masses (d-f) ROI without masses

GREY-LEVEL DIFFERENCE STATISTICS

The Gray level difference method is based on the histogram of absolute difference between pairs of gray level. Several texture measure like mean, standard deviation, entropy; contrast can be obtained from Gray level Difference Statistics. This is shown in [Fig. 3].

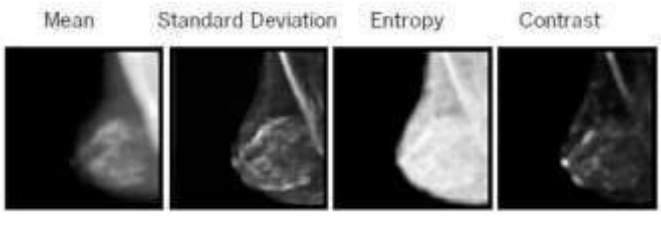


Fig. 3: Texture measure from Gray level difference statistics.

FLOW DIAGRAM

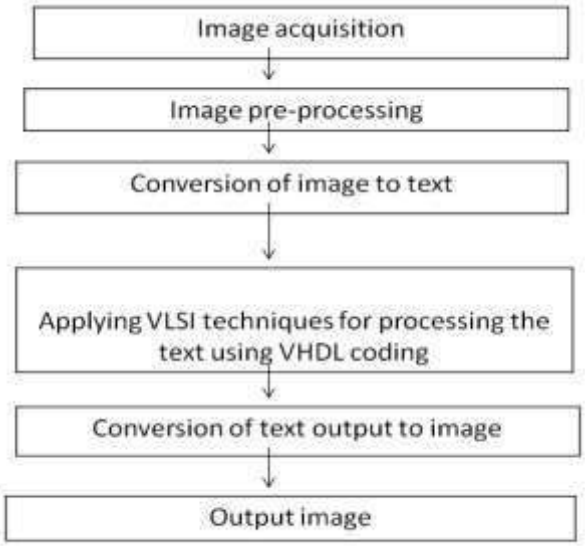


Fig. 4: Block Diagram of the proposed system

IMAGE ACQUISITION

In [Fig. 4] ,Image acquisition in image processing can be broadly defined as the action of retrieving an image from some source, usually a hardware-based source, so it can be passed through whatever processes need to occur afterward. Performing image acquisition in image processing is always the first step in the workflow sequence because, without an image, no processing is possible. The image that is acquired is completely unprocessed and is the result of whatever hardware was used to generate it, which can be very important in some fields to have a consistent baseline from which to work. One of the ultimate goals of this process is to have a source of input that operates within such controlled and measured guidelines that the same image can, if necessary, be nearly perfectly.

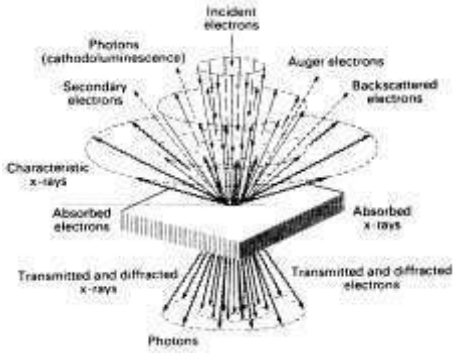


Fig. 5: signals generated by the interaction of an electron beam with a solid reproduced under the same conditions so anomalous factors are easier to locate and eliminate.

IMAGE PRE-PROCESSING

Image pre-processing can significantly increase the reliability of an optical inspection. Several filter operations which intensify or reduce certain image details enable an easier or faster evaluation. Users are able to optimize a camera image with just a few clicks.

CONVERSION OF IMAGE TO TEXT FILE

An image is a 2D matrix. For converting into text file, it needs to be converted into 1D matrix and then to text file using file write in matlab. The image size is also a constraint, so it needs to be fixed for conversion. Then converted to 1D matrix using reshape command. Then write into text file.

```
A=imread('1.bmp');
B=rgb2gray(A);
disp('Image file read successful');
%Fig.ure,imshow(B),title  ('org');
C=imresize(B,[isize isize]);
Fig.ure,imshow(C),title('croped');
d=reshape(C,1,[]);
disp('Reshapping done');
fid = fopen('img.txt', 'wt');
fprintf(fid, '%8d\n', d);
disp('Text file write done');disp(‘ ‘);
fclose(fid);
```

TESTER MODULE
APPLYING VLSI TECHNIQUES

This module tests the flip-flop by generating the clock and D signal of the timing diagram above and dumping the Q and QN signals of the flip-flop. It's outputs are the flip-flop's inputs and viceversa.

```
//Tester module sends a periodic clock signal to the flip-flop
module tester(q,qn,clk,d);
input q,qn;
output clk,d;
reg clk,d;

//Run the test once
initial
begin
        clk=0;
        //Dump results of the simulation to ff.cvd
        $dumpfile("dff.vcd");
        $dumpvars;
        //Generate input signal d
         d=0; #9 d=1; #1 d=0; #1 d=1; #2 d=0; #1 d=1; #12 d=0;
        #1 d=1; #2 d=0; #1 d=1; #1 d=0; #1 d=1; #1 d=0; # 7 d=1;
        #8 $finish;
end

//Generate periodic clock signal
always
begin
        #4 clk=!clk;
end
endmodule
```

This module is behavioral too as we have initial and always blocks. You should be able to understand most of the code. However, there are a few new concepts here. The $dumpfile and $dumpvars commands tell the Verilog simulator (more on this ahead) to log the module's variables to the specified file, "dff.vcd" in this case. You may also be wondering what the #s do. These are Verilog delays. The delay the following instruction by a given amount of time. For example, #4 clk=!clk; within an always block changes "clk" every four time units from 0 to 1, producing a square wave. The time unit is a second by default. Without using delays there is no way of making the program work. This is the way to control time in the design. You may add delays to any instruction. For example, you could model the flip-flop's delay by adding some to its always block. It's now easy to understand how the d=0; #9 d=1; #1 d=0; #1 d=1; ... lines produce the D signal we want. Finally, the $finish command tells the simulator to stop the simulation once the D signal was generated. If this command was omitted the simulation would continue indefinetly because this time the always block has no condition (there is no @ like in the flip-flop module).

TEST BENCH MODULE

This module just connects the tester module to the flip-flop module:

```
//Test bench connects the flip-flop to the tester module
module testbench;
        wire clk,d,q,qn;
        dff ff1(d,clk,q,qn);
        tester tst1(q,qn,clk,d);
endmodule
```

It is the most simple of the modules, but it's very different. This time it's structural code, that is, you define the structure of the circuit. It's like describing the circuit diagram. In this case the final circuit is simply the flip-flop connected to the tester. To create a flip flop use dff ff1(d,clk,q,qn);. First goes the module name, followed by the part name, which could be almost any string, followed by the wires that connect to the module in parenthesis. These must follow the order in the module's declaration. In a structural module we use wires. Regs are not necessary because they are defined inside the different modules.

COMPILING AND SIMULATING

Go ahead and copy/paste the modules into a text file, order doesn't matter. Call the file "dff.v". The .v extension is standard for Verilog files, but isn't required by the compiler. To compile open a Command Prompt at your working directory (where you saved the file). A quick way to open a command prompt at any directory is to hold shift and right-click the folder, then click "Open Command Window Here". Type:
iverilog -o dff dff.v
The "-o" tells Icarus Verilog to "save output to the following file". The output is then saved to "dff". This file is not executable. It has to be run using vvp, the Icarus Verilog simulator which is the one that actually produces simulation results, zeros and ones for each of the model variables, as a function of time. To run the simulation type:

vvp dff

This is what outputs the dff.vcd file with all the simulation results. However if you open this file with your text editor you'll see it's not easy to understand. To generate an easy-to-understand timing diagram from this file we use GTKWave.

GTKWave does have a GUI. To open it press Windows key and type "gtkwave". Then click File -> Open New Tab and chose the ffd.vcd file. Now you must add the variables in order to see their timing diagram. Click on "testbench" at the left (SST panel) and then select all the variables using Ctrl or Shift and "Insert" them.

If everything is okay you should get a timing diagram exactly as the one at the beginning of the tutorial, just like the following:

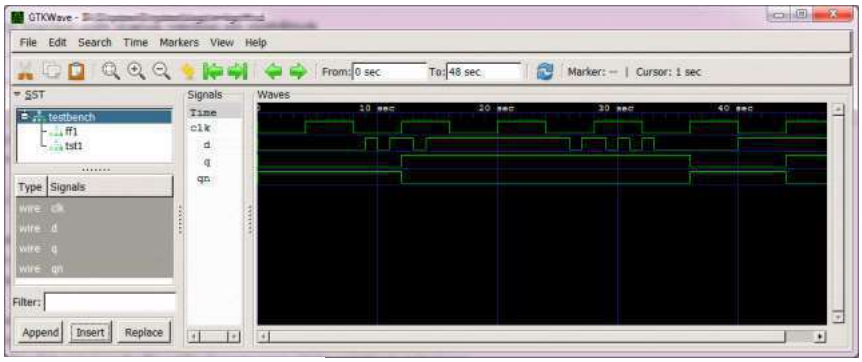


Fig. 6: Output Image of VLSI

When testing your programs you'll have to go to the compiling-simulating-loading process every few minutes. Remember you can use the up-down arrow keys while in the command prompt to access the last commands and compile/simulate. On GTKWave use File->Reload Waveform to reload the vcd file and refresh the timing diagrams without having to reload each variable. By using these tips the whole process will take you a few seconds. It's over. Now feel free to change the code around to see what happens. Mastering the use of delays, wires and rags takes some time. See Verilog_in_One_Day for a more in depth explanation of the language shown in [Fig. 6].

CONVERTING TEXT FILE INTO IMAGE

For converting text file into image, it needs to be converted from 1D matrix and then to image. The image size must same as previous for processing text into image. Then converted to 1D matrix using reshape command. Then write into text file. First, the text file is read using file read as vector (1D matrix). Then it has to be converted to matrix. Finally transpose to complete it as image.

```
fidh = fopen('img.txt');
Ah = fscanf(fidh, '%g %g', [1 inf]);
disp('Text file read done');
fclose(fidh);
S1h= vec2mat(Ah,isize,isize);
disp('Vector conversion done');
%c=imresize(S1,[128 128]);
Sh= transpose(S1h);
Jh=uint8(Sh);
disp('Image is ready');
imwrite(Jh,['newimage','.jpg']);
Fig.ure,imshow(Jh),title('IMAGE form TEXT file');
```

OUTPUT

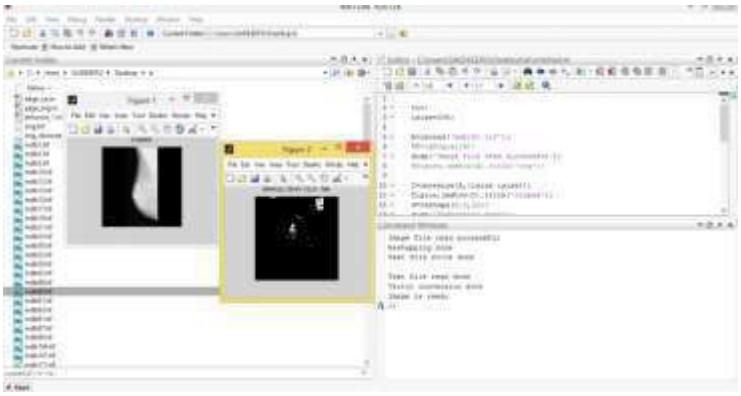


Fig. 7: Image Data base

IMAGE DATABASE

The methods were tested with a total of 56 images (each of size 1024 1024 pixels at a resolution of 200 m) including 30 benign breast masses, 13 malignant, and 13 normal cases selected from the mini MIAS, database. The dataset includes circumscribed and spiculated cases in both benign and malignant categories. The cases in each category (benign and malignant) are further subdivided in the database as dense-glandular, fatty, and fatty-glandular types. Table I shows the numbers of various types of cases in each category. The mean values of the sizes of the masses are 1.07+_0.77 cm And 1.22+_0.85 cm for the benign and malignant categories, respectively. The radius of the smallest mass (malignant) is 0.34 cm, and that of the largest mass (benign) is 3.9 cm. The center of abnormality and an approximate radius of each mass are indicated in the database. The circular demarcation of masses as done in the database is not useful for confirming the results of mass detection as such a demarcation may also include normal fibro-glandular tissues in the ROIs, particularly in spiculated cases. Hence, in the present work, the center of abnormality as indicated for each mass in the database was used to confirm the result of mass detection. Karssemeijer and to Brake, and Petrick etc. used a similar criterion for confirming the results of their mass detection algorithms.Karssemeijer and te Brake detected spiculated malignant tumors and lesions corresponding to architectural distortion using 19 malignant cases present in the MIAS database. Additionally, in the present work, the efficacy of the mass regions segmented in terms of benign versus malignant discrimination shown in [Fig. 7].

RESULTS

IMAGE TO TEXT CONVERSION

Initially the input is read on MATLAB. Then the image is cropped as show in the below shown in [Fig. 8].

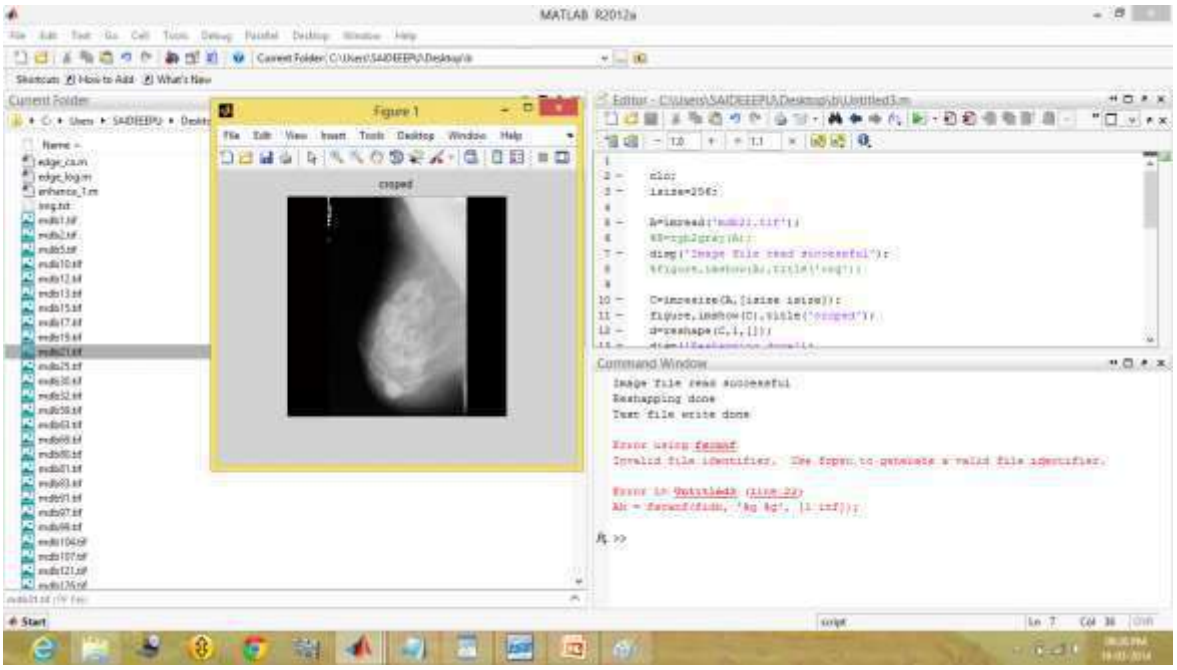


Fig. 8: Image to Text Conversion

IMAGE TO TEXT FILE

After reading the image, it is further converted into a text file and the intensity values of the text file is as below .shown in [Fig. 9].

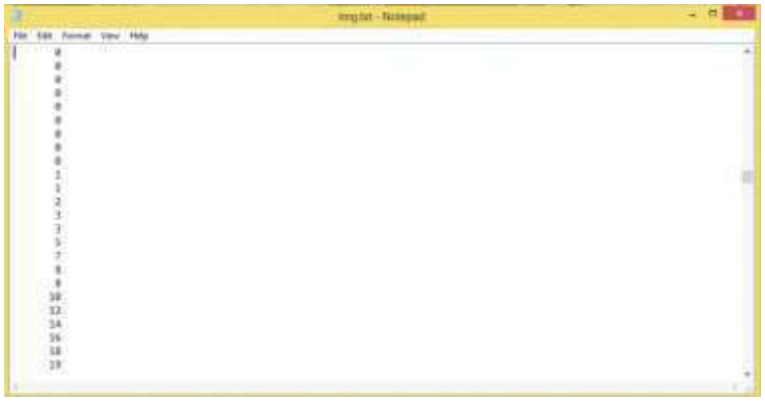


Fig. 9: Image To Text File

XILINX INPUT (TEXT)

The converted image is then given as input in the form of text on Xilinx and VERILOG is developed for the classification of cancer cells present in the image as shown in [Fig. 10] using Xilinx software.

Fig. 10: XILINX Text

XILINX OUTPUT (WAVE FORMS)

The output for the classification of cancer cells in the text file is obtained in the form of wave forms on Xilinx as shown in [Fig.11]

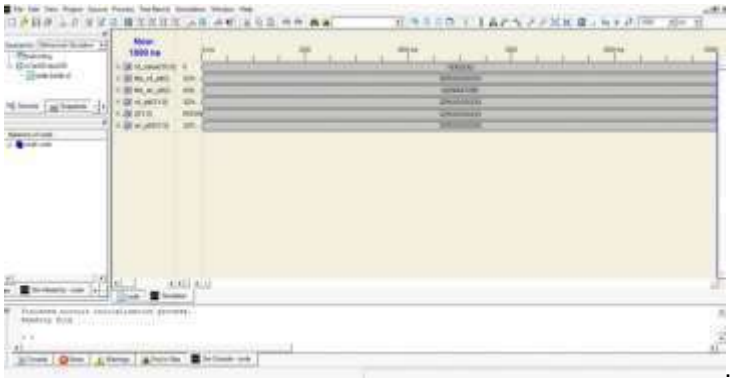


Fig. 11: XILINX Output

TEXT FILE OUTPUT INTO IMAGE

A text file is created according to the output waveforms on Xilinx. This text file is then read on MATLAB and converted into image in which cancer cells are classified as shown in [Fig. 12].

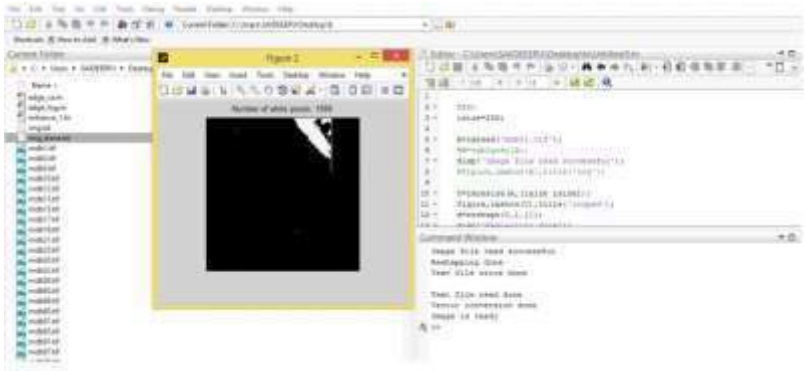


Fig. 12: Text File Output in to image

NUMBER OF CANCER CELLS AFFECTED

Finally after classification of cancer cells, the number of affected cells is calculated in MALAB and the result is shown in [Fig. 13].

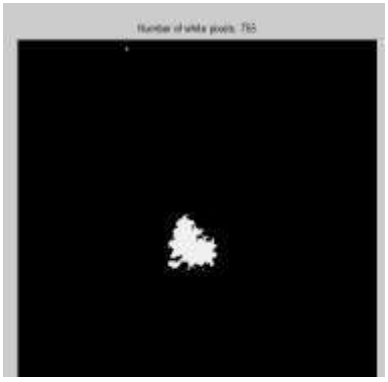


Fig.13: Number of cancer cells affected in the image.

CONCLUSION

Breast cancer is one of the major causes of death among women. Digital mam- mography screening programs can enable early detection and diagnose of the breast cancer which reduces the mortality and increases the chances of complete programs produce a great amount of mammographic images which have to be interpreted by radiologists. Due to the wide range of breast ab- normalities' features some abnormalities may be missed or misinterpreted. There is also a number of false positive findings and therefore a lot of unnecessary biop- sies. Computer-aided detection and diagnosis algorithms have been developed to help radiologists give an accurate diagnosis and to reduce the number of false positives. There are a lot of algorithms developed for detection of masses and cal- cifications. In this chapter, algorithms that are commonly used and the ones re- cently developed were presented. Over the years there has been an improvement in the detection algorithms but their performance is still not perfect. The area under the ROC curve is rarely above 90% which means that there are still many false positive outputs. Possible reason for such a performance may be the characteristics of breast abnormalities. Masses and calcifications are sometimes superimposed and hidden in the dense tissue which makes the segmentation of correct regions of interest difficult. Another issue is extracting and selecting appropriate features that will give the best classification results. Furthermore, the choice of a classifier has a great influence on the final result and classifying abnormalities as benign or ma- lignant is a difficult task even for expert radiologists. Further developments in each algorithm step are required to improve the overall performance of computer-aided detection and diagnosis algorithms

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] AntoBennet M, Sankar Babu G, Natarajan S, [2015] Reverse Room Techniques for Irreversible Data Hiding , Journal of Chemical and Pharmaceutical Sciences 08(03): 469-475.

[2] AntoBennet, M , Sankaranarayanan S, Sankar Babu G, [2015] Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis, Journal of Chemical and Pharmaceutical Sciences 08(03): 571-576.
AntoBennet, M, Suresh R, Mohamed Sulaiman S, [2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface, Journal of Chemical and Pharmaceutical Research 07(08): 291-299.

[3] AntoBennet, M & JacobRaglend, [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, Journal of Computer Science, 8(09): 1447-1454.

[4] AntoBennet, M & JacobRaglend, [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, European Journal of Scientific Research, 66(1):120-129.

[5] AntoBennet, M , Resmi R. Nair, Mahalakshmi V, Janakiraman G [2016] Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography, Indian Journal of Science and Technology, 09 (02):01-07.

[6] AntoBennet, M & JacobRaglend, [2012] A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, Journal of Applied Sciences Research, 8(5) :2429-2438.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

PERFORMANCE & ANALYSIS OF HYPER SPECTRAL IMAGE COMPRESSION USING FAST DISCRETE CURVELET TRANSFORM WITH ENTROPYCODING

M Anto Bennet[1*], A Ahamed Meeran Mydeen[2], Kelwin Inasu[3], C Suthesh[4], M Venkatesh[5]
[1]Professor,[2]Asst.professor, [3,4,5]UG Students, Department of ECE, VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

The work presents the efficient hyper spectral image compression using fast discrete curve let transform and AAC(Adaptive Arithmetic Coding) and comparison with SPECK coding. Image compression is the technique uses lossy or lossless coding to reduce the storage space required for image information by removing spatial redundancy. Lifting wavelet transform based set partitioning in embedded block coder is used for compression. Along with this system discrete curve let based adaptive arithmetic entropy coding algorithm will be used for effective compression to increase compression ratio with minimum error rather than lossy coding. SPECK coder performs compression based on priority based transmission with two sets partitioning such as LIS and LSP. Fast discrete curve let transform is used to decompose the retargeted image into set of coefficients called approximation and detailed one in different orientations. The detailed coefficients contains both noise and edge information. The high frequency sub bands represents edges and redundant information extracted from all curved regions. I n encoding stage, Adaptive arithmetic coding is involved to shrink the coefficients contains to represent the image in minimal number of bits. Then bit stream will be transmitted or stored and compression ratio will be measured. At the decoder side, bit streams are decoded and then perform inverse fast discrete curve let transformation for reconstructing an image. This system evaluates the performance of AAC coding with various bit rates in terms of processing time, mean square error and correlation. The simulated results will be shown that used algorithm has lower complexity with high performance in terms of CR and image construction rather than lossy set partitioning in embedded block coder.

KEY WORDS
Adaptive Arithmetic Coding(AAC), Discrete Cosine Transform (DCT)

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email :
bennetmab@gmail.com

INTRODUCTION

A considerable measure of exploration has been done on nursery agro frameworks and all the more for the most part on secured harvests to control vermin and ailments by natural means rather than pesticides. Research in horticulture is pointed towards increment of efficiency and nourishment quality at decreased use and with expanded benefit, which has gotten significance in late time. A solid request now exists in numerous nations for non-concoction control techniques for vermin or illnesses. Nurseries are considered as biophysical frameworks with inputs, yields and control process circles. A large portion of these control circles are automatized (e.g., atmosphere and fertirrigation control).The administration of lasting organic product crops requires close checking particularly for the administration of illnesses that can influence generation altogether and consequently the post-harvest life. In the event of plant the illness is characterized as any disability of ordinary physiological capacity of plants, delivering trademark manifestations.

An indication is a wonder going with something and is viewed as proof of its presence. Sickness is brought about by pathogen which is any operators bringing on malady. In the greater part of the cases vermin or illnesses are seen on the leaves or stems of the plant. Thusly distinguishing proof of plants, leaves, stems and discovering the bug or infections, rate of the irritation or sickness frequency, manifestations of the nuisance or illness assault, assumes a key part in effective development of products. In natural science, in some cases a huge number of pictures are created in a solitary examination. These pictures can be required for further studies like characterizing injury, scoring quantitative characteristics, ascertaining territory eaten by creepy crawlies, and so forth. All of these assignments are prepared physically or with particular programming bundles. It is colossal measure of work as well as experiences two noteworthy issues: extreme handling time and subjectiveness ascending from various people [5, 6, 7, 8].

MATERIALS AND METHODS

Efficient hyper spectral image compression based on Fast Discrete Curve let Transform with Adaptive Arithmetic Coding and performance analysis. The performance parameters are Compression Ratio (CR), Mean Square Error (MSE), Peak Signal to Noise Ratio(PSNR), Correlation and Elapsed Time. The input is the RGB image. It is a additive colour image in which the three colours are added together to produce various colours shown in [Fig. 1]. [Fig. 2] shows the hyper spectral input image. The main purpose of RGB image is sensing, representation and display of images. Here the input image used is hyper spectral image. The size of compressed hyper spectral image is 256×256.

Editor: Dr. K. Sakthisudhan

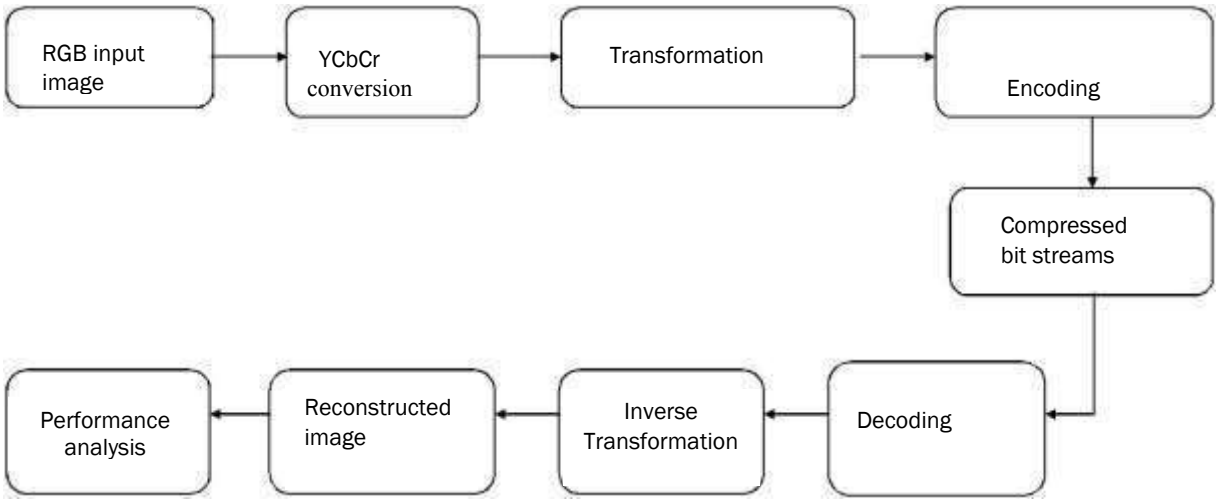


Fig. 1: Various colours sshow the hyper spectral input image

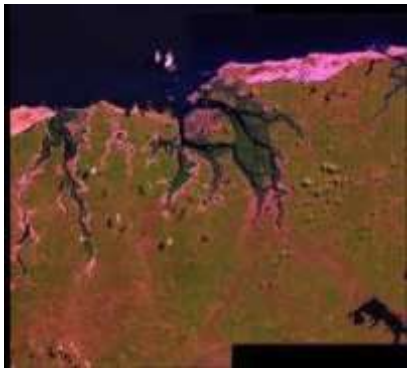


Fig. 2: Various colours show the hyper spectral input image

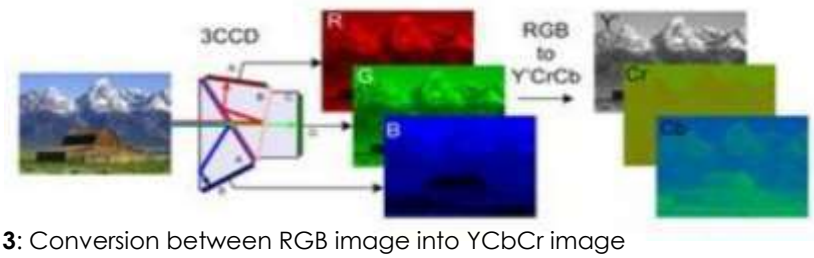


Fig. 3: Conversion between RGB image into YCbCr image

The YCbCr conversion represent the colour conversion from RGB to yCbCr.It is the family of colour images used as a part of colour image pipe line in video and digital photography systems.It is not an absolute colour space rather it is a way of encoding technique. [Fig 3] shows the conversion between RGB image into YCbCr image. YCbCr signals are called YPbPr , and are created from the corresponding gamma-adjusted RGB (Red, Green, Blue) source using two defined constants KB and KR.

ADAPTIVE ARITHMETIC ENTROPY CODING

Arithmetic coding is different from other coding methods for which have the exact relationship between the coded symbols and the actual bits that are written to a file. It codes one data symbol at a time, and assigns to each symbol a real-valued number of bits. To figure out how this is possible, we have to understand the code value representation: coded messages mapped to real numbers in the interval [0, 1]. The code value v of a compressed data sequence is the real number with fractional digits equal to the sequence's symbols. Convert sequences to code values by simply adding \0." to the beginning of a coded sequence, and then interpreting the result as a number in base-D notation, where D is the number of symbols in the coded sequence alphabet. For example, if a coding method generates the sequence of bits 0011000101100, then it have

Code sequence d = [0011000101100)]

Code value v = 0.0011000101100 2 = 0.19287109375

Where the '2' subscript denotes base-2 notation. As usual, omit the subscript for decimal notation. This construction creates a convenient mapping between infinite sequences of symbols from a D-symbol alphabet and real numbers in the interval [0, 1), where any data sequence can be represented by a real number, and vice-versa. The code value representation can be used for any coding system and it provides a universal way to represent large amounts of information independently of the set of symbols used for coding (binary, ternary, decimal, etc.). For instance, in [1.5] the same code with base-2 and base-10 representations. Evaluate the of any compression method by analyzing the distribution of the code values it produces. From Shannon's information theory, if a coding method is optimal, then the cumulative distribution of its code values has to be a straight line from point [0, 0] to point [1, 1].It introduces the notation and equations that describe arithmetic encoding, followed by a detailed example. Fundamentally, the arithmetic encoding process consists of creating a sequence of nested intervals.

Coding algorithm

Arithmetic coding assigns a sequence of bits to a message, a sting of symbols. Arithmetic coding can treat the whole symbols in a list or in a message as one unit. Unlike Huffman coding, arithmetic coding doesn´t use a discrete number of bits for each. The number of bits used to encode each symbol varies according to the probability assigned to that symbol. Low probability symbols use many bit, high probability symbols use fewer bits. The main idea behind Arithmetic coding is to assign each symbol an interval. Starting with the interval [0...1], each interval is divided in several subinterval, which its sizes are proportional to the current probability of the corresponding symbols. The subinterval from the coded symbol is then taken as the interval for the next symbol. The output is the interval of the last symbol [1, 3].In order to clarify the arithmetic coding; we explain the previous example using this algorithm. [Table 1] depicts the probability and the range of the probability of the symbols between [0] and [1].

The [table 1] represents the input message consists of the following symbols: 2 0 0 136 0 and it start from left to right. [Fig. 3.5] depicts the graphical explanation of the arithmetic algorithm of this message from left to right.

Table 1: Table of probability and ranges distribution of symbols

Symbols	Probability	Range
0	0.63	[0,0.63)
2	0.11	[0.63,0.74)
14	0.1	[0.74,0.84)
136	0.1	[0.84,0.94)
222	0.06	[0.94,1.0)

RESULTS

Fig. 4: Input image

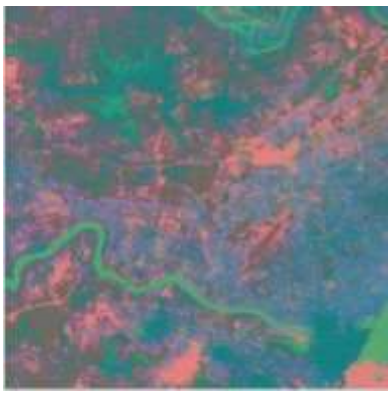


Fig. 5:Color Space Conversion

The [Fig. 4] represents the input hyper spectral image .The size of an image is in the range of 1390*836 .In this range cannot get better output image and the compression time is also more , so it is compressed into 300*300 to reduce the compression time .Therefore the required output image will be obtained. Color Space Conversion is a preprocessing module .The image should be represented in matrix form that is

mxnx3 form so the images are classified into three layers. The three layers are red, green, and blue (RGB colors).In order to make the processing simple the yCbCr color conversion is used.In The YCbCr color space conversion, y represents the luminance yellow, Cb represents chromium blue and Cr represents the chromium red .The luminance yellow gives information about the brightness ,structural and the texture of the images .The chrominance blue and red gives color information.

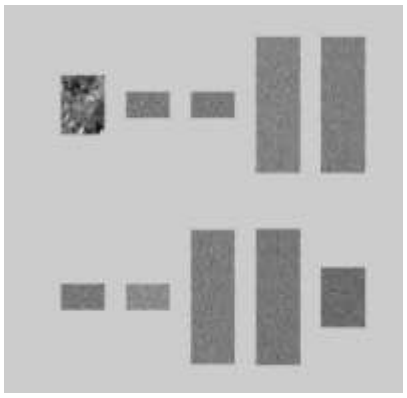


Fig. 6.FDCT decomposition

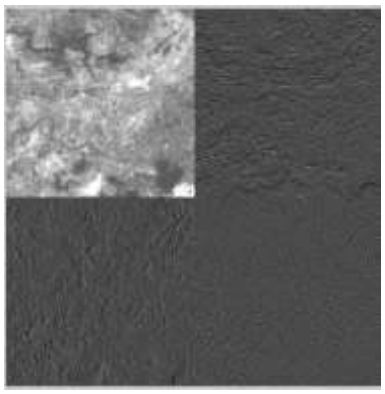


Fig. 7: LWT Decomposition.

The Fast Discrete Curve let transform [fig. 6] has three steps. The three steps are Wavelet transform, FFT transform and Ridge let transform. The chromium blue and chromium red which specifies the color information such as repeated information, irrelevant compression ratio, redundant data of an image undergoes plane separation. The FDCT transform has 9 high frequencies and 1 low frequency. The diagram [Fig. 7] represents 8 different orientation. It is the representation of images in different scales in different orientations shown in [Fig. 7]. of the image shown in [Fig. 5].

Fig.7: i) Using SPIHT

Fig.7: ii) Using AAC

Here is the comparison between the SPIHT coding output and AAC coding output [Fig 8].Comparing to SPIHT coding Adaptive Arithmetic Coding has Less Distortion and better compression ratio. Flexibility of adaptive Arithmetic Coding is higher than SPIHT coding and there is no need of any code books at decoder side. It gives high compatibility shown in [Fig 9-12] Also measures in corresponding [Tables 1-4].

PERFORMANCE MEASURES OF COMPRESSION RATIO ANALYSIS

Table 1: Compression ratio Analysis for SPIHT vs AAC coding.

Image Name	Using SPIHT	Using AAC
Ocean image	2.1013	12.0656
Highway Area	1.7724	3.6772
Multitemporal image	1.3986	2.5607
Chennai Beach	2.0203	6.3853
Urban Area	1.6490	4.4292
Multispectral Image	2.1422	7.8990

Waubay	1.9833	5.9263
Lake	1.3380	2.2430
Agriculture Land	1.9062	5.2845

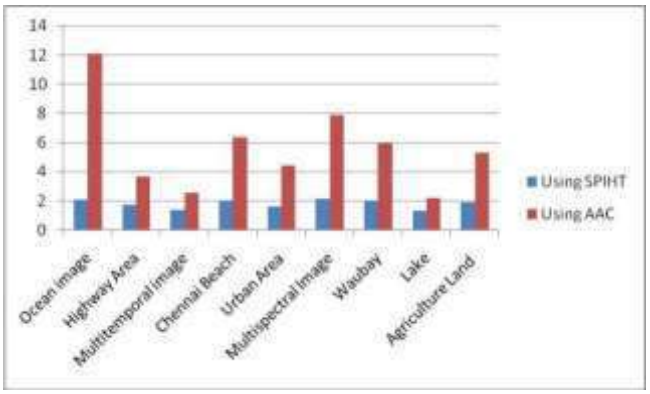


Fig. 9: Compression ratio Analysis

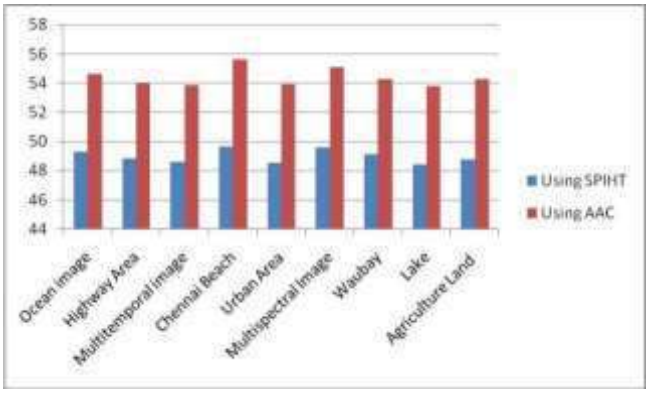


Fig. 10: Mean Square Error Analysis

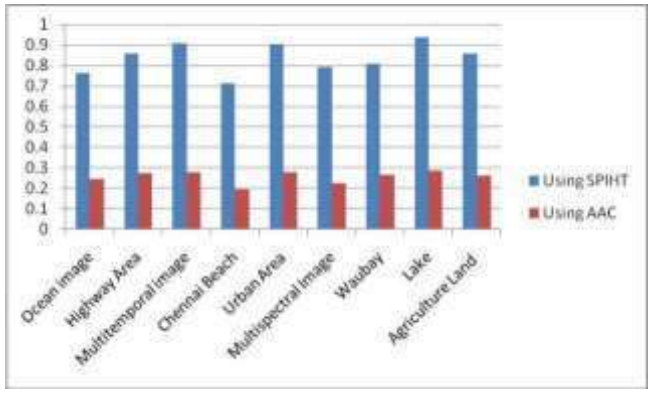


Fig. 11: Peak Signal To Noise Ratio analysis

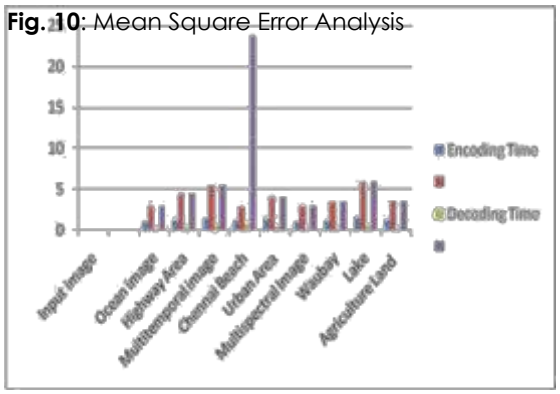


Fig. 12: Encoding time and Decoding Time Analysis

SQUARE ERROR ANALYSIS

Table 2: Mean Square Error Analysis for SPIHT vs AAC coding

Input image	Using SPIHT	Using AAC
Ocean image	0.7633	0.2461
Highway Area	0.8565	0.2754
Multitemporal image	0.90919	0.2800
Chennai Beach	0.7123	0.1944
Urban Area	0.9058	0.2802
Multispectral Image	0.7928	0.2236
Waubay	0.8068	0.2647
Lake	0.9386	0.2869
Agriculture Land	0.8563	0.2612

Peak Signal To Noise Ratio Analysis

Table 3: Peak Signal To Noise Ratio analysis for SPIHT vs AAC coding

Input image	Using SPIHT	Using AAC
Ocean image	49.3265	54.6216
Highway Area	48.8143	53.9990
Multitemporal image	48.5795	53.8776
Chennai Beach	49.6817	55.6318
Urban Area	48.5757	53.8925
Multispectral Image	49.6095	55.0831
Waubay	49.1304	54.2511
Lake	48.4250	53.7908
Agriculture Land	48.8114	54.2794

Encoding time and Decoding Time Analysis

Table 4: Encoding time and Decoding Time Analysis for SPIHT vs AAC coding

Input image	Encoding Time		Decoding Time	
	Using SPIHT	Using AAC	Using SPIHT	Using AAC
Ocean image	1.0832	2.9643	0.2747	2.9275
Highway Area	1.25236	4.4485	0.4217	4.5043
Multitemporal image	1.5082	5.4502	0.6564	5.5228
Chennai Beach	1.1351	2.7737	0.851	23.7949
Urban Area	1.3970	4.0470	0.5062	4.1113
Multispectral Image	1.0491	2.9206	0.2310	2.9162
Waubay	1.1780	3.4120	0.3526	3.4663
Lake	1.5585	5.8842	0.6553	5.9255
Agriculture Land	1.2037	3.5318	0.3769	3.5317

CONCLUSION

This work presented that HS image compression using lossy and lossless coding for analysis of compression effect based on wavelet based set partitioning in hierarchical trees coding and curve let based arithmetic coding. Wavelet based speck coding was used to compress image by considering significant coefficients based on priority. Adaptive arithmetic coding provided better encoding performance with discrete curve let transform. It represented an image interms of detailed coefficients in all directions. The simulated results shows that, entropy coding provides better compression ratio rather than SPIHT coding and Image quality also can preserve it with entropy lossless Coding. Further the hyper spectral image compression by using fast discrete curve let transform with entropy coding system will be enhanced by modifying used encoding algorithm with context adaptive coding to preserve the image details and with low complexity.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Mohammad H Asghari and Bahram Jalali, [2014] Discrete Anamorphic Transform For Image Compression-IEEE signal procesing letters, 21(07): 829-833.

[2] Chenwei Deng, Weisi Lin, and Jianfei Cai., [2012] Content-Based Image Compression for Arbitrary-Resolution Display Devices , IEEE transactionson multimedia, 14(04): 1127-1139.

[3] Jong-Woo Han, Kang-Sun Choi, Tae-Shick Wang, Sung-Hyun Cheon, and Sung-Jea Ko, [2009] Improved Carving Using a Modified Energy Function Based on Wavelet Decomposition- The 13th IEEE International Symposium on Consumer Electronics .

[4] Emmanuel J Candès and Michael B Wakin, [2008] An Introduction To Compressive Sampling [A sensing/sampling paradigm that goes against the common knowledge in data acquisition] IEEE SIGNAL PROCESSING MAGAZINE .21-30.

[5] Liron Yatziv and Guillermo 6Sapiro, [2006] Fast and Video Colorization Using Chrominance Blending - IEEE TRANSACTIONS ON IMAGE PROCESSING, 15(0 5): 1-15.

[6] AntoBennet, M, Sankar Babu G, Natarajan S, [2015] Reverse Room Techniques for Irreversible Data Hiding, Journal of Chemical and Pharmaceutical Sciences 08(03): 469-475.

[7] AntoBennet, M , Sankaranarayanan S, Sankar Babu G, [2015] Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis, Journal of Chemical and Pharmaceutical Sciences 08(03): 571-576.

[8] AntoBennet, M, Suresh R, Mohamed Sulaiman S, [2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface, Journal of Chemical and Pharmaceutical Research 07(08): 291-299.

[9] AntoBennet, M [2015] A Novel Effective Refined Histogram For Supervised Texure Classification, International Journal of Computer & Modern Technology , Issue 01 (02): 67-73.

[10] AntoBennet, M, Srinath R,Raisha Banu A, [2015]Development of Deblocking Architectures for block artifact reduction in videos, International Journal of Applied Engineering Research, 10: 6985-6991.

[11] AntoBennet, M & JacobRaglend, [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, Journal of Computer Science, 8.(09): 1447-1454.

[12] AntoBennet, M & JacobRaglend, [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, European Journal of Scientific Research, 66(01):120-129 .

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

IMPROVEMENT OF AUTOMATIC HEMORRHAGES DETECTION METHODS IN RETINAL IMAGING AND IMAGE ANALYSIS

M Anto Bennet[1*], J Surekha Poomathi[2], C Kalpana[3], S Sariga Priya[4]

[1]Professor,[2],[3,4,]UG Students,Department of ECE, VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

Our vision reduced in eye due to the presence of Retinal diseases like Exudates (diabetic Retinopathy), Micro aneurysms, and Blood vessel damage. This work mainly concentrates on the symptoms of heart, lung, liver and kidney problems identification using Retinal fundus images. Our proposed work shows that how optic disk elimination and follower the symptom detection. Optic disk is one of the parts which consist of intersection of blood vessels and it also has same characteristics of exudates like yellow color, intensity and contrast. Distinguish the exudates and optic disk is critical one. So only first eliminate the optic disk and follower that exudates detection .This detection method very favorably with existing and promise deployment of these systems. Micro aneurysms are the initial stage of exudates.

KEY WORDS
Adaptive Histogram Equalization (AHE), Gray-Level Co-occurrence Matrix (GLCM), Contrast limited AHE (CLAHE).

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

Hemorrhage is defined as an escape of blood from ruptured blood vessel. Ischemia is a term used to describe a tissue whose blood supply has been reduced to an insufficient level. Lack of 02 in the retinal tissue may lead to retinal cell death and result in reduced vision shown in [Fig. 1].

The problems in kidney and Lung are detected through Micro aneurysm formation in Retina. The kidneys are bean shaped organs that serve several essential regulatory roles in vertebrate animals [1]. They are essential in the urinary system and also serve homeostatic functions such as the regulation of electrolytes, maintenance of acid–base balance, and regulation of blood pressure (via maintaining salt and water balance). They serve the body as a natural filter of the blood, and remove wastes, which are diverted to the urinary bladder. In producing urine, the kidneys excrete wastes such as urea and ammonium, and they are also responsible for the reabsorption of water, glucose, and amino acids. The kidneys also produce hormones including calcitriol, erythropoietin, and the enzyme rennin. Located at the rear of the abdominal cavity in the retroperitoneal, the kidneys receive blood from the paired renal arteries, and drain into the paired renal veins. Each kidney excretes urine into a urethra, itself a paired structure that empties into the urinary bladder.Renal physiology is the study of kidney function, while nephrology is the medical specialty concerned with kidney diseases. Diseases of the kidney are diverse, but individuals with kidney disease frequently display characteristic clinical features. Common clinical conditions involving the kidney include the nephritic and nephrotic syndromes, renal cysts, acute kidney injury, chronic kidney disease, urinary tract infection, nephrolithiasis, and urinary tract obstruction. Various cancers of the kidney exist; the most common adult renal cancer is renal cell carcinoma [2,3]. Cancers, cysts, and some other renal conditions can be managed with removal of the kidney, or nephrectomy. When renal function, measured by lomerular filtration rate, is persistently poor, dialysis and kidney transplantationmay be treatment options. Although they are not normally harmful, kidney stones can be painful and repeated, chronic formation of stones can scar the kidneys. The removal of kidney stones involves ultrasound treatment to break up the stones into smaller pieces, which are then passed through the urinary tract. One common symptom of kidney stones is a sharp to disabling pain in the medial/lateral segments of the lower back or groin shown in [Fig. 2].

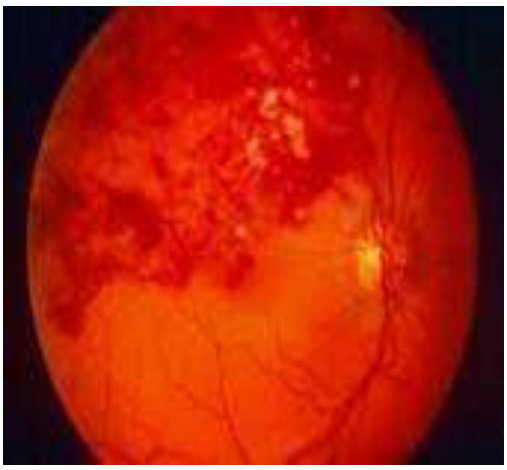


Fig. 1: Retina with hemorrhage formation

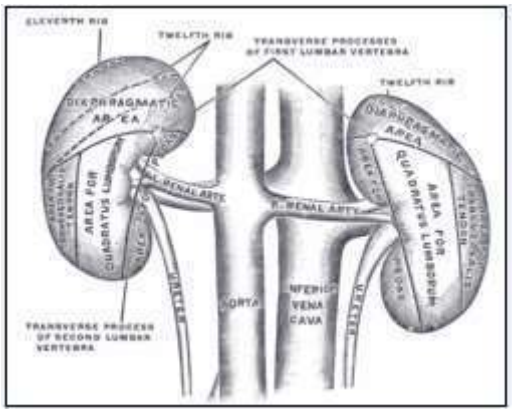


Fig. 2: Diagram of Kidney

3D-rendered computed tomography, showing renal arteries and veins. The kidneys receive blood from the renal arteries, left and right, which branch directly from the abdominal aorta. Despite their relatively small size, the kidneys receive approximately 20% of the cardiac output [4].Each renal artery branches into segmental arteries, dividing further into interlobar arteries, which penetrate the renal capsule and extend through the renal columns between the renal pyramids. The interlobar arteries then supply blood to the arcuate arteries that run through the boundary of the cortex and the medulla. Each arcuate artery supplies several interlobular arteries that feed into the afferent arterioles that supply the glomeruli. The medullary interstitium is the functional space in the kidney beneath the individual filters (glomeruli), which are rich in blood vessels [5,6]. The interstitium absorbs fluid recovered from urine. Various conditions can lead to scarring and congestion of this area, which can cause kidney dysfunction and failure.After filtration occurs the blood moves through a small network of venules that converge into interlobular veins. As with the arteriole distribution the veins follow the same pattern, the interlobular provide blood to the arcuate veins then back to the interlobar veins, which come to form the renal vein exiting the kidney for transfusion for blood. The kidney participates in whole-body homeostasis, regulating acid-base balance, electrolyte concentrations, extracellular fluid volume, and regulation of blood pressure. The kidney accomplishes these homeostatic functions both independently and in concert with other organs, particularly those of the endocrine system. Various endocrine hormones coordinate these endocrine functions; these include renin, angiotensin II, aldosterone, antidiuretic hormone, and atrial natriuretic peptide, among others.Many of the kidney's functions are accomplished by relatively simple mechanisms of filtration, reabsorption, and secretion, which take place in the nephron [7]. Filtration, which takes place at the renal corpuscle, is the process by which cells and large proteins are filtered from the blood to make an ultrafiltrate that eventually becomes urine. The kidney generates 180 liters of filtrate a day, while reabsorbing a large percentage, allowing for the generation of only approximately 2liters of urine. Reabsorption is the transport of molecules from this ultrafiltrate and into the blood. Secretion is the reverse process, in which molecules are transported in the opposite direction, from the blood into the urine.Although the kidney cannot directly sense blood, long-term regulation of blood pressure predominantly depends upon the kidney. This primarily occurs through maintenance of the extracellular fluid compartment, the size of which depends on the plasma sodium concentration. Renin is the first in a series of important chemical messengers that make up the renin-angiotensin system [8]. Changes in renin ultimately alter the output of this system, principally the hormones angiotensin II and aldosterone. Each hormone acts via multiple mechanisms, but both increase the kidney's absorption of sodium chloride, thereby expanding the extracellular fluid compartment and raising blood pressure. When renin levels are elevated, the concentrations of angiotensin II and aldosterone increase, leading to increased sodium chloride reabsorption, expansion of the extracellular fluid compartment, and an increase in blood pressure. Conversely, when renin levels are low, angiotensin II and aldosterone levels decrease, contracting the extracellular fluid compartment, and decreasing blood pressure [9].

MATERIALS AND METHODS

The Retinal fundus images are converted to either green component or gray scale for feature extraction. The green channel extractions are then fed into image enhancement. Low contrast images could often due to several reasons, such as poor (or) non uniform lightning condition, non-linearity (or) small dynamic range of image sensor. Image enhancement features are then fed into morphological dilation, erosion and opening operation. Dilation is used for expanding an element. Erosion is used for shrinking an element. And opening operation is used to remove a optical disk using radius formula. The circular border is then fed into pixel classification. In pixel classification uses a techniques namely Graycoprops. Graycoprops normalizes the gray-level co-occurrence matrix (GLCM) so that the sum of its elements is equal to 1.There are four techniques are used to classify the GLCM are contrast, correlation,homogeneity and energy. The

input image operation is performed using mat lab syntax. Finally, the output images (i.e.,) Blood Vessels, micro-aneurysm and exudates are obtained shown in [Fig. 3].

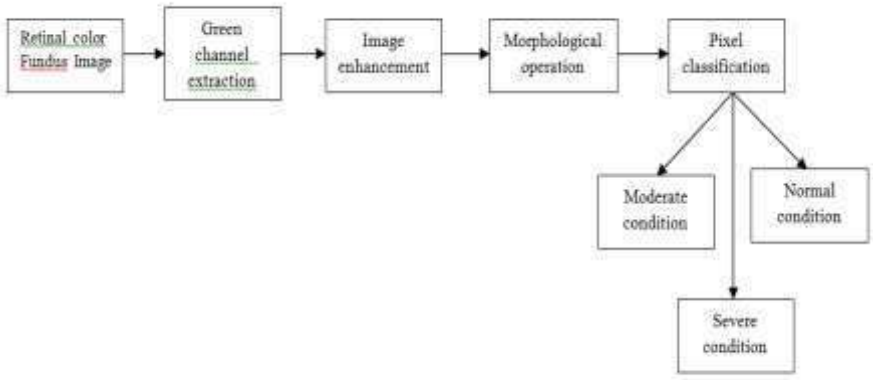


Fig. 3: Block diagram of automatic hemorrhages detection method

RETINAL COLOR FUNDUS IMAGES

The color retinal images are taken using fundus camera. It consists of three colors red, green, and blue. Since background of retina is red in color we omit red color in the image and blue color have high wavelength of noise and it is omitted and we take only green channel in the image shown in [Fig. 4].

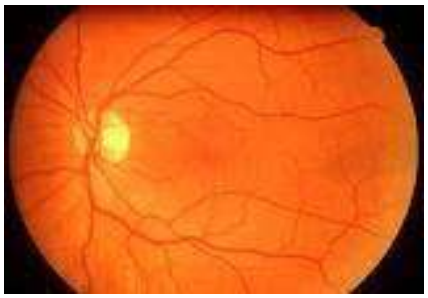


Fig. 4: Color fundus image

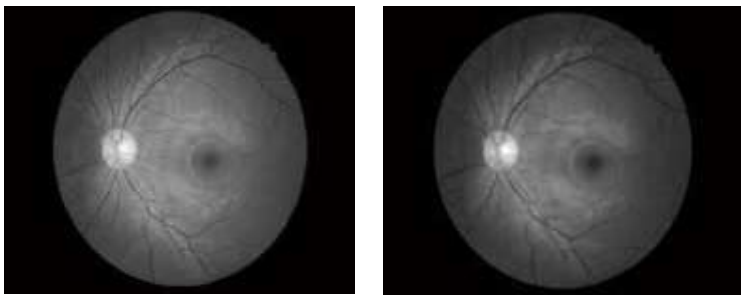


Fig. 5: Green channel extraction and gray scale image.

GREEN CHANNEL EXTRACTION

The color fundus images are converted to either green component or grayscale for features extraction of texture analysis. Green channel provides maximum contrast between background and foreground shown in [Fig. 5].

IMAGE ENHANCEMENT

The normalized features values are then fed into image enhancement. Low contrast images could often due to several reasons, such as poor (or) non uniform lightning condition, non-linearity (or) small dynamic range of image sensor.

ADAPTIVE HISTOGRAM EQUALIZATION

Adaptive Histogram Equalization (AHE) is a computer image processing technique used to improve contrast in images. It differs from ordinary histogram equalization in the respect that the adaptive method computes several histograms, each corresponding to a distinct section of the image, and uses them to redistribute the lightness values of the image. It is therefore suitable for improving the local contrast of an image and bringing out more detail. However, AHE has a tendency to over amplify noise in relatively homogeneous regions of an image shown in [Fig. 6].

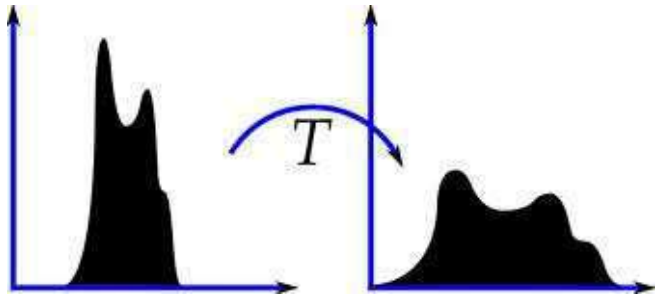


Fig. 6: Adaptive Histogram Equalization

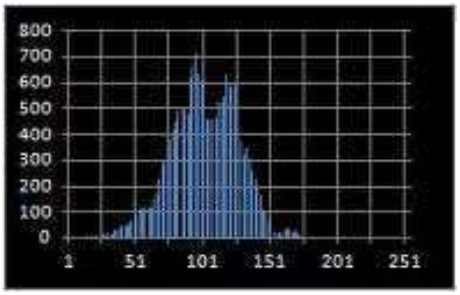


Fig. 7: Real Time Performance of AHE

Experience has shown that the size of the contextual region can be set to either 1/16 or 1/64 of the image area for essentially all medical images, with the smaller region chosen only when the feature size of interest is quite small. With a smaller contextual region the contrast becomes too sensitive to very local variations and in particular to image noise. The oversensitivity to local variations can cause artifacts, which have never been experienced with the preferred contextual region sizes. Although AHE frequently produces excellent result in certain cases noise becomes disturbingly obvious. In particular, this occurs where the image includes relatively homogeneous regions or a poor signal to noise ratio [Fig. 7]. Contrast limited AHE (CLAHE) avoids this over enhancement of noise. Contrast enhancement can be defined as the slope of the function mapping input intensity to output intensity. We will assume that the range of input and output intensities is the same. Then a slope of 1 involves no enhancement, and higher slopes give increasingly higher enhancement. Thus the limitation of contrast enhancement can be taken to involve restricting the slope of the mapping function. With histogram equalization the mapping function m (i) is proportional to the cumulative Histogram:

m (i) =(Display-Range) • (Cumulative_Histogram(i)/ Region-Size) (1)

Since the derivative of the cumulative histogram is the histogram, the slope of the mapping function at any input intensity, i.e. the contrast enhancement, is proportional to the height of the histogram at that intensity: dm/di =(Display-Range/Region- Size) .histogram(i).Therefore limiting the slope of the mapping function is equivalent to clipping the height of the histogram shown in [Fig. 8].

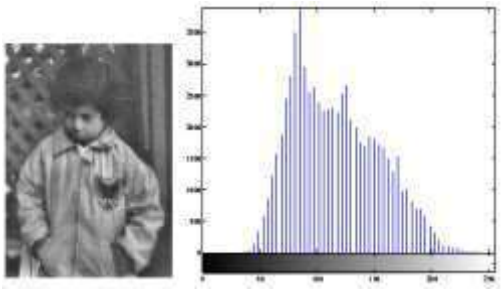


Fig. 8: Image after CLAHE equalization with its histogram .

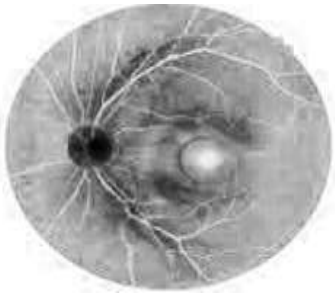


Fig. 9: Image and histogram after adaptive histogram equalization

CLAHE operates on small regions in the image, called tiles, rather than the entire image. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches the histogram specified by the 'Distribution' parameter. The neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image shown in [Fig. 9].

MORPHOLOGICAL OPERATION

Based on shapes input size is equal to the output size. The each pixel in output image is compared to corresponding pixel in the input image with its neighbors.

Detection of retinal vessels

The input retinal images are first to be detected for abnormal vessels in the retinal region. Studies show that earlier techniques such as the holding and histogram equalization have shown cases of missing out thin vessels which is a serious problem. The main reason behind is the resultant image obtained via the techniques implemented previously were blurred with lack of accuracy. The proposed methodology aimed at performing the curve let transform for effective enhancement of the image for better viewing along with effective edge detection using the multi structure morphological reconstruction for detecting the edges along eight directionalities unlike mathematical morphology

Dilation

Dilation is used for expanding an element A by using structuring element B Dilation of A by B and is defined by the following equation:

$$A \oplus B = \{z \mid (B)_z \cap A \neq \phi\} \quad (2)$$

This equation is based on obtaining the reflection Of B about its origin and shifting this reflection by z. The dilation of A by B is the set of all displacements z, such that and A overlap by at least one element.

Based On this interpretation the equation of can be rewritten as:

$$A \oplus B = \{Z \mid [(B)Z \cap A] \subset A\} \quad (3)$$

Erosion

Erosion is used for shrinking of element A by using element B.Erosion for Sets A and B in Z2, is defined by the Following equation:

$$A \oplus B = \{z \mid [(B)z \cap A] \subset A\} \quad (4)$$

This equation indicates that the erosion of A by B is the set of all points z such that B, translated by z, is combined in A shown in [fig 10&11].

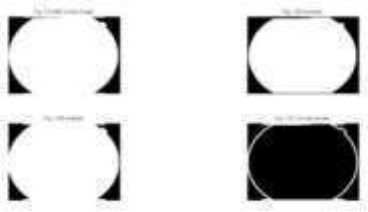


Fig. 10: Output of erosion and dilation of an image

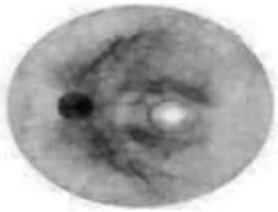


Fig.11: Image after imerode and imdilate

FEATURE EXTRACTION

Different features of the fundus images namely Blood vessels, Exudates and Micro aneurysms are extracted using image processing techniques. The values obtained are essential as they represent the image and are necessary in order to classify the images accurately.

EXPERIMENTAL PROCEDURE FOR BORDER FORMATION

There are two methods in detecting the circular border of an image. Both methods are essential as each method could not work for few of the images due to their contrast intensity. Deploying both methods allow the detection of all the images. Border formation is to clean off the noisy edges and is also use during exudates, micro aneurysm and blood vessel.

Border formation method 1

Grayscale image instead of green channel is used as it more efficient for border formation. The first method uses canny method to detect the edges before enclosing the circular region with a top and bottom bar. Function "imfill" is then applied to fill the region. The circular border is then formed by subtracting the eroded and dilated images shown in [Fig 12].

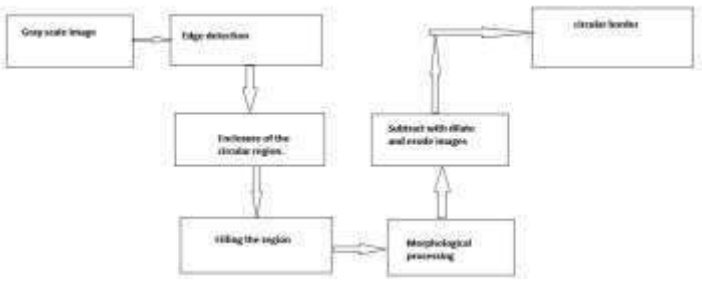


Fig. 12: border formation method 1

Border formation method 2

Method 2 is activated when a noisy image is obtained instead of a circular border. This method inverses the intensity of the images first before image segmentation is applied with function. The circular region is filled as a result and the circular border obtained after subtracting the dilated image with eroded image shown in [Fig. 13 &14].

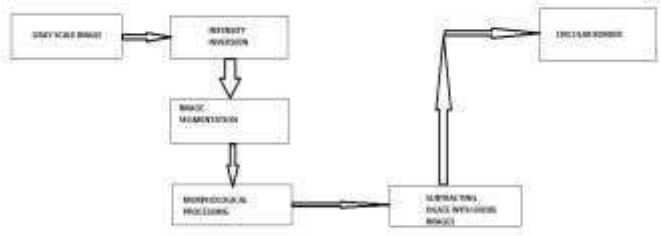


Fig. 13: Border formation method 2

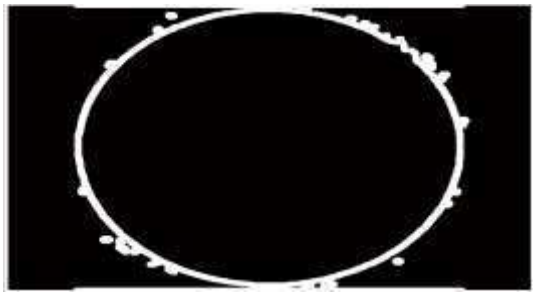


Fig.14: Subtracted image.

EXPERIMENTAL PROCEDURE-MASK OPERATION FOR OPTIC DISK

As optic disk is made up of a group of bright spots, it is not suitable to use loops and locate the largest value. This would only point to one spot and most likely to be on the side of optical disk. The mask required to cover optical disk would be inefficient as it would be much larger and covers more details. Mask creation is used in the detection of blood vessels, exudates and micro aneurysm. After locating the optic disk, a mask needed to be created. A simple mask created using loops would be easy but it would be result in error when the optic disk is close to the image shown in [Fig. 15& 16].

R2=(x-k)2+(y-h)2 (5)

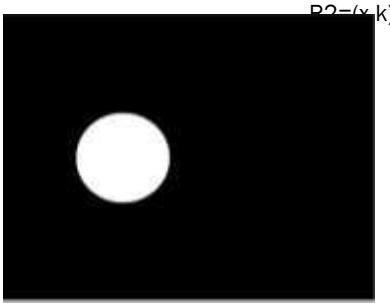


Fig. 15: Removal of optic disk.

Fig. 16: Image after removal of optic disk

EXPERIMENTAL PROCEDURE-AND LOGIC

Two methods of detecting blood vessels are used. Both methods generally detect different location of the images like exudates as blood vessels; hence by computing their similarity the non blood vessels area could be filtered. AND logic is applied to mark out the similar pixels of the two images. The output pixel is registered as binary1 (white) when both images' pixels are binary1(white) the obtained image would be clearer shown in [Fig. 17&18].

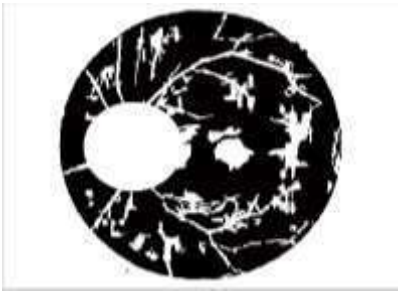


Fig.17: Blood vessel with Noise & mask

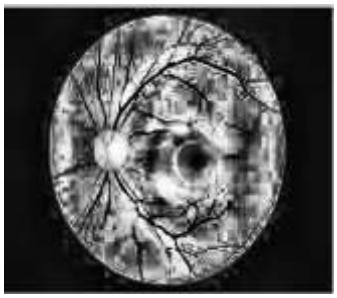


Fig.18: Blood vessel &noise after adaptive histogram

RESULT

The area of blood vessel is obtained using two loops to count the number of pixels with binary 1(white) in the final blood vessel image shown in [Fig.19].Finally the corresponding values are shown in [Table 1].

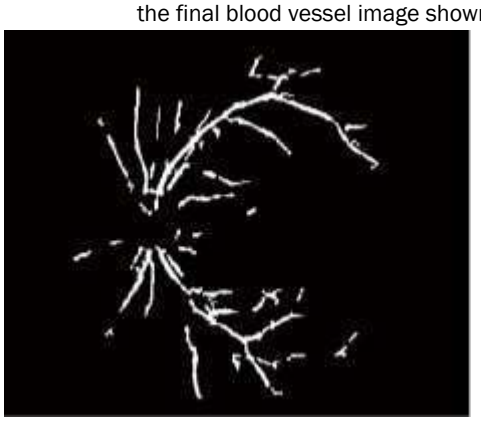


Fig.19: Final blood vessel image

Editor: Dr. K. Sakthisudhan

Table 1. Threshold Values For Blood Vessels

AREA	ENERGY	CONTRAST	CORRELATION	HOMOGENITY	OUTPUT
15435	0.269534	0.941772713	0.121151769	0.871993035	0
14176	0.270413	0.936932181	0.123631065	0.871076244	0
17575	0.300596	0.932113324	0.115210245	0.860918304	0
9638	0.260851	0.933853832	0.127305377	0.873249401	1
32667	0.299253	0.959811786	0.091914368	0.867089627	0
34750	0.304575	0.951999122	0.099191053	0.863451239	0
21092	0.31617	0.936177946	0.108731471	0.858249425	0
28889	0.327915	0.949214444	0.09618857	0.856598872	0
18964	0.303315	0.938694965	0.110899929	0.861114411	0
9301	0.273335	0.927134808	0.1276279	0.873631998	1
28758	0.362864	0.944921501	0.091267655	0.842319056	0
8521	0.263922	0.932088982	0.125012336	0.872845918	1
9475	0.274166	0.916474163	0.147605483	0.868018733	1
14917	0.279434	0.926004474	0.128673513	0.867605913	0
25961	0.340012	0.949237017	0.090816972	0.861026881	0
22105	0.346957	0.926848875	0.108737955	0.846607098	0

CONCLUSION

Biomedical image processing requires an integrated knowledge in mathematics, statistics, programming and Biology. The values obtained are essential as they represent the image and are necessary in order to classify the images accurately. Based on the result of the classifier, this project has a sensitivity of 80% and specificity of 20%. It is able to achieve a fairly accurate classification for mild and higher stages, but not for normal class resulting in a possible high false alarm. This might be improved by fine tuning the threshold values used on the images and more images could be used to improve the overall system. In this paper, we learnt various techniques of image processing and were able to extract the features namely blood vessels, exudates and micro aneurysms and texture properties like area, energy, contrast, correlation and homogeneity from the fondues images.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] CJ Vyborny, ML Giger, and RM Nishikawa. [2000] Computer aided detection and diagnosis of breast cancer," Radiologic Clinics N. Amer., 38(4):725–740, 10.1016/S0033-8389(05)70197-4, 0033-8389.

[2] J Scharcanski and CR Jung, [2006] Denoising and enhancing digital mammographic-Images for visual screening, Computerized Medical Image, Graphics, 30 (4):243–254, 10.1016/j. comp med image, 2006.05.002, 0895-6111

[3] AntoBennet, M, Sankar Babu G, Natarajan S, [2015] Reverse Room Techniques for Irreversible Data Hiding, Journal of Chemical and Pharmaceutical Sciences 08(03): 469-475.

[4] AntoBennet M, Sankaranarayanan S, Sankar Babu G,[2015] Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis, Journal of Chemical and Pharmaceutical Sciences 08(03): 571-576.

[5] Dr. AntoBennet, M, Suresh R, Mohamed Sulaiman S,[2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", Journal of Chemical and Pharmaceutical Research 07(08): 291-299.

[6] AntoBennet, M & JacobRaglend, [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, Journal of Computer Science, 08(09): 1447-1454.

[7] AntoBennet, M & JacobRaglend, [2001] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, European Journal of Scientific Research, 66(01): 120-129.

[8] AntoBennet M, Resmi R Nair, Mahalakshmi V, Janakiraman G [2016] Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography, Indian Journal of Science and Technology, 09(02):01-07.

[9] AntoBennet M & JacobRaglend. [2012] A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, Journal of Applied Sciences Research,8(05): 2429-2438.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

ADAPTIVE FINGER PRINT IMAGE ENHANCEMENT WITH EMPHASIS

M Anto Bennet[1*], R Srinath[2], D Abirami[3], S Thilagavathi[4], S Soundarya[5]

[1]Professor,[2]Asst.professor,[3,4,5]UG Students,Department of ECE, VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

This work proposes several improvements to an adaptive fingerprint enhancement method that is based on contextual filtering. The term adaptive implies that parameters of the method are automatically adjusted based on the input fingerprint image. Five processing blocks comprise the adaptive fingerprint enhancement method. Hence, the proposed overall system is novel. The four updated processing blocks are: 1) preprocessing 2) global analysis 3) local analysis 4) matched filtering. In the preprocessing and local analysis blocks, a nonlinear dynamic range adjustment method is used. In the global analysis and matched filtering blocks, different forms of order statistical filters are applied. These processing blocks yield an improved and new adaptive fingerprint image processing method. The performance of the updated processing blocks is presented in the evaluation part of this paper. The algorithm is evaluated toward the NIST developed NBIS software for fingerprint recognition on FVC databases

KEY WORDS
Region of Interest (ROI), Finger Print , atching(FPM).,Histogram Equalization(HE)

INTRODUCTION

Digital image processing is currently a hot research in image enhancement and it believed that they will receive extensive application to security purpose in the next few years. Finger print image is the pattern of ridges and valleys also called furrows in the finger print literature on the surface of a finger trip. Each individual has unique finger print. The uniqueness of a fingerprint is exclusively determined by the local ridge characteristics and their relationships [1,2,3]. A total of 150 different local ridge characteristics like Islands, short ridges, enclosure, etc., have been identified. Image pre-processing is the most evaluative step for accurate minutiae detection and fingerprint matching. Accurate estimation of overall steps of the algorithm is very important for reliable result. In this research paper, minutiae based fingerprint matching technique is studied in detail and implemented in MATLAB.

This research paper shows analyzer can recognize the fingerprint image by minutiae point calculation as well as location evaluation of minutiae points. This method has successfully been applied to the generation of synthetic fingerprints of the same finger. However, it is hard to estimate the parameters accurately due to insufficient information. In local minutiae matching, these approaches can be considered as an effective tool [4, 5]. However, as the size of the tolerance boxes had to be increased, the probability of falsely matching fingerprints from different fingers also increases. Some methods used local similarity measures to improve the robustness of the distortions since fingerprint images are less affected by distortions in the local area. Many fingerprint matching methods have been developed to cope with distortions , most of them are minutiae-based. Thus, they cannot use more topological information (such as ridge shape) covering the entire fingerprint image and the limitation of information still exists Existing methods typically keep various parameters, such as local area size, constant. The strategy to keep parameters constant may fail in a real application where fingerprint image or sensor characteristics vary, thus yielding varying image quality. In addition, due to the spatially variable nature of fingerprints, it is crucial to have a sufficient amount of data in each local image area so that the local structure of the fingerprint is enclosed. Hence, the local area size should adapt to data present. Different fingerprint sensor resolutions provide different normalized spatial frequencies of the same fingerprint and this also requires adaptive parameters [6, 7, 8].

PROPOSED SYSTEM

A spatial sinusoidal signal and its corresponding magnitude spectrum is illustrated together with a local finger print image patch and its corresponding magnitude spectrum in [Fig. 1]. The ridge feature vectors between the minutiae in the ridge coordinate system can be expressed as directional graph whose nodes are minutiae and whose edges are ridge feature vectors. Thus, we can adopt graph matching methods to utilize the ridge feature vectors in fingerprint matching. They first defined the local neighborhood of each minutia, called K-plet, which consists of the K–nearest minutiae from a center minutia. The comparison of two K–plets is performed by computing the distance between the two strings obtained by concatenating the neighboring minutiae, sorted by their radial distance with respect to the center minutia. Neighborhoods are matched by dynamic programming and a match of local neighborhoods is propagated with a breadth-first fashion. Thus, we apply this matching scheme to our ridge-based coordinate system, since the ridge-based coordinate system can be represented as a graph and each coordinate system makes a local neighborhood [9, 10].

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

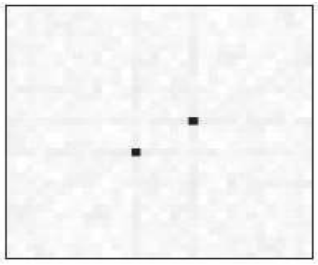


Fig. 1(a): Sinusoidal signal

Fig. 1(b): Magnitude spectrum of sinusoidal image

Fig. 1(c): Local area of inside finger print

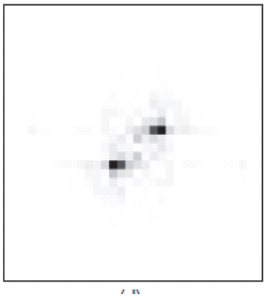


Fig. 1 (d): Magnitudinal signal between sinusoidal signal and fingerprint

They first defined the local neighborhood of each minutia, called K-plet, which consists of the K-nearest minutiae from a center minutia. The comparison of two K-plets is performed by computing the distance between the two strings obtained by concatenating the neighboring minutiae, sorted by their radial distance with respect to the center minutia.

MATERIALS AND METHODS

Methodology

Existing method keep various parameter such as local area size, constant. The strategy to keep parameters constant may fail in a real application where fingerprint image or sensor characteristics vary, thus yielding varying image quality. In addition due to the spatially variable nature of fingerprints, it is crucial to have a sufficient amount of data in each local image area so that the local structure of the fingerprint is enclosed. Hence, the local image area size should adapt to the data present. Different

fingerprint sensor resolutions provide different normalized spatial frequencies of the same fingerprint and this also requires adaptive parameters. Fingerprints captured with the same sensor may also vary depending on, e.g., gender and age of the user in [Fig. 2] and [Fig. 3]The negative influence on fingerprint recognition system performance for individuals of different ages was demonstrated and the matching results of Db3 in FVC2000

Fig. 2: Fingerprint sensor image of the little finger of 30 years old man.

Fig. 3: Fingerprint image of 5 years old boy

To compensate for varying fingerprint image characteristics and to achieve an optimal system performance, key parameters of most existing methods, e.g., the size of the local area, need to be tuned manually for every fingerprint image. This manual tuning for each image is tedious and costly and automatic systems are therefore desirable.

Steps involved in proposed system

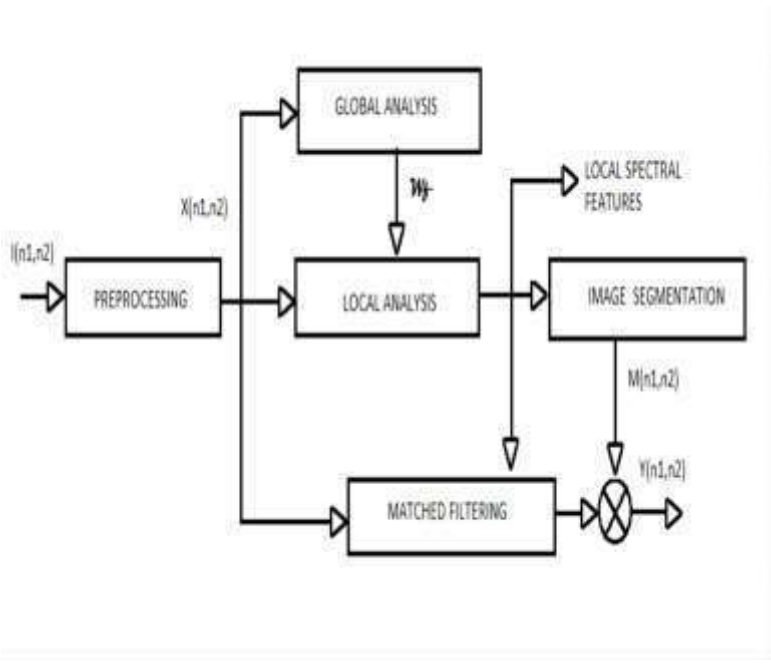


Fig. 4: Block diagram of proposed system

Preprocessing

I(n1,n2) represents a fingerprint image of size N1*N2,where n1 1nd n2 denotes horizontal and vertical coordinates. Each element of image is quantized in 256 gray scale levels, i.e., the dynamic range of the image is adjusted by using SMQT (Successive Mean Quantization Transform shown in [Fig. 4]).

Global analysis

Novel update to the previously derived global finger print analysis is conducted to aid the fundamental spatial frequency estimation of the finger print image and improves the frequency performance for noisy images.

Local adaptive analysis

The purpose of the local analysis is to adaptively estimate local features corresponding to fingerprint ridge frequency and orientation. Most parts of a fingerprint image containing ridges and valleys have, on a scale, similarities to a sinusoidal signal in noise. Hence, they have a magnitude spectrum with two distinct spectral peaks located at the signals dominant spatial frequency and oriented in alignment with the spatial signal.

Matched filtering

It is based on the spectral features estimated in the local analysis, where an addition order statistical filtering of the spectral features is introduced to increase the method's resilience towards noise. It adjusts the fundamental frequency to match the local image area and improve spectral features estimation.

Fig. 5: Images of different minutiae.

An illustration of the fingerprint enhancement with variant 1 and variant 16 of the proposed method is given in [Fig. 5].The most pronounced visual effect is that fingerprints processed with the new method preserve larger parts of the original fingerprint, parts which were excluded in the original method. Another observation is that scars in the fingerprint are better removed (especially in the rightmost example), and this is attributed to a combination of the improved feature extraction and the added order statistical filtering of the matched filter design.

RESULTS

The restored fingerprint images will be more suitable than the original images for visual examination and/or automatic feature examination. The fingerprint image is first normalized to reduce the variations of gray-level values along the ridges and valleys, the orientation fields are computed based on chain code.

The region of interest are then segmented from background using the method described by Ratha, the segmented fingerprint images are filtered by the composite filter, the images can be binarized adaptively, finally the ridge contour following algorithm is utilized to extract endings and bifurcations minutiae, and filtering performance and efficiency are evaluated correspondingly. In the following two subsections, the specific methods in this paper for orientation computation and binarization are explained below.

Get finger print image

In this first Module, store the fingerprint image [Fig. 6] in a folder, to get and display that image by using push button.

Estimate the quality and segment the image

The typical feature procedures as well as additional procedures for quality estimation and circular variance estimation the mean and variance values of each block are calculated to segment the fingerprint regions in the image shown in [Fig. 7].

Estimate the orientation

Orientation calculation is critical for fingerprint image enhancement and restoration in both frequency and spatial domain. Without exception, the computation of the orientation image in the proposed algorithm will affect directly the enhancement efficiency. In the current literature, most of the fingerprint classification and identification processes calculate the local ridge orientation of fixed-size block instead of each pixel. The most popular approach is based on binary image gradients, other approaches have been proposed in different research groups. An innovative computational method, based on chain code is proposed in our lab.

Chain code is loss less representation of gray-level image in terms image recovery. The chain code representations of fingerprint image edges capture not only boundary pixel information, but also the counter-clockwise ordering of those pixels in the edge contours. Therefore, it is convenient to calculate direction for each boundary pixel. In our calculation, end points and singular points, which are detected by the ridge flow following method, are not used for computation with chain code with chain code elements less than 20 are regarded as noises and excluded for orientation computations shown in [Fig. 8].

Estimate the frequency

To define the sign of the vertical axis according to the origin, the cross product between the orientation of the origin and the vector pointing from the origin to the side of the vertical axis shown in [Fig. 9].

Extract the minutiae

The quality estimation is performed to avoid extracting false minutiae from poor quality regions and to enhance the confidence level of extracted minutiae set shown in [Fig. 10].

Skeletonization

The Gabor filter is applied to enhance the image and obtain a skeleton zed ridge image. Then, the minutiae (end points and bifurcations) are detected in the skeleton zed image shown in [Fig. 11].

Binarization

Robust preprocessing method also used to reduce enhancement errors. Moreover, ridge features can be used in other applications. In the area of fingerprint identification, it is important to be able to extract alignment-free features since it needs no time to align a query feature set with the enrolled feature sets one by one.

The fingerprint images possess ridge flow patterns slowly changes in directions. They may have various gray level values due to non-uniformity of the ink intensity, non-uniform contact with the sensors by users or changes in illumination and contrast during image acquisition process. Global threshold method fails to create good quality binary images for further feature extractions. In Greenberg's work, adaptive thresholding is used to binarize fingerprint images, binarization depends on the comparison result of gray-level value of each pixel with local mean. Adaptive binarization method based on Clustering of background and foreground pixels, i.e., Otsu algorithm. Otsu method selects the optimal threshold b minimizing the within-class variance of the groups of pixels separated by the thresholding operator shown in [Fig. 12].

Estimate the circular variance

The orientation estimation process, the performance can be improved. Second, there are some minutiae pairs offering no ridge feature vectors because some images had small foreground regions or their levels of quality were too low, as can be seen in the foreground region was very small and there were few minutiae shown in [Fig. 13].

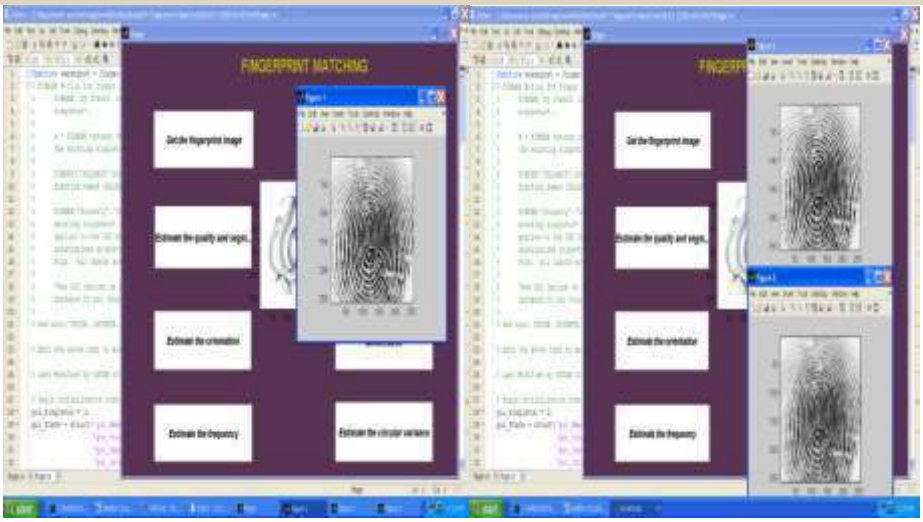


Fig. 6: Fingerprint image

Fig. 7: Segmentation of image

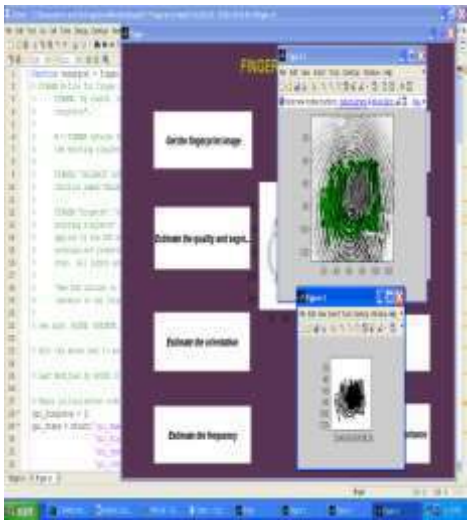


Fig. 8: Orientation of image

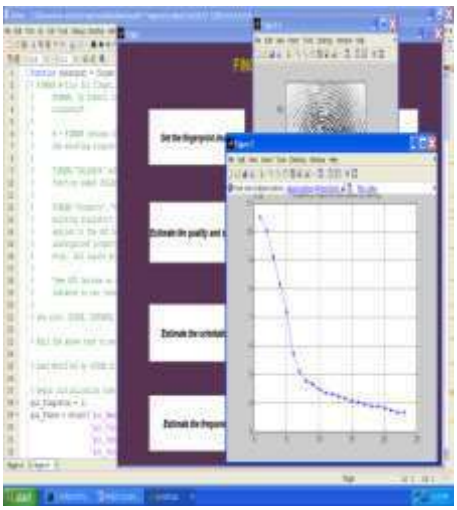


Fig. 9: Frequency estimation of image

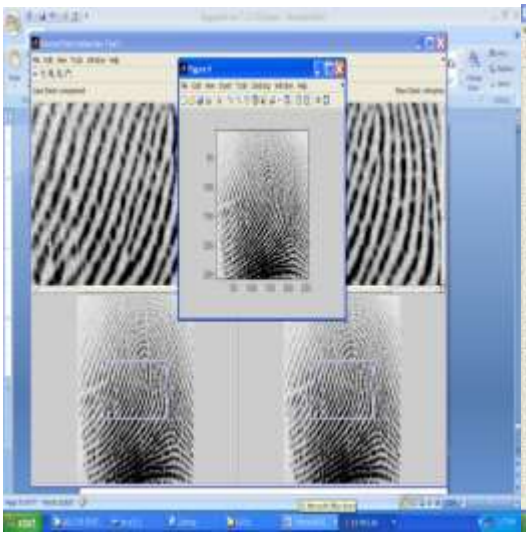


Fig. 10: Extract the minutiae

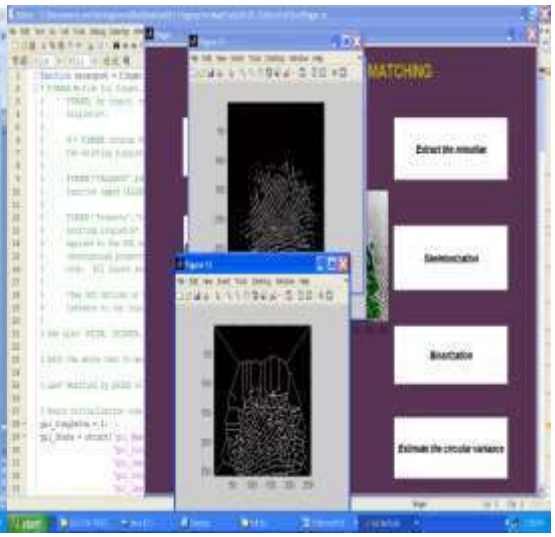


Fig. 11: Image of skeletonization

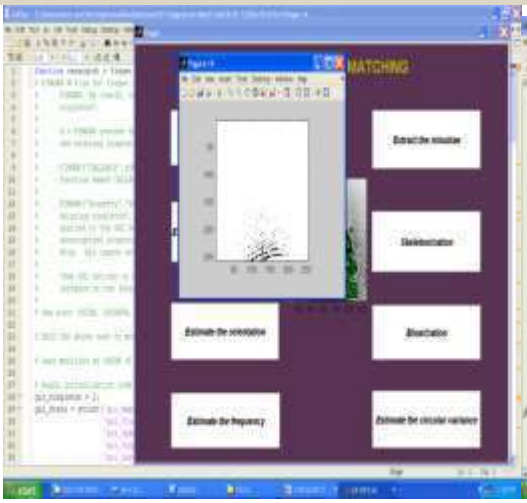


Fig. 12: Image of binarization.

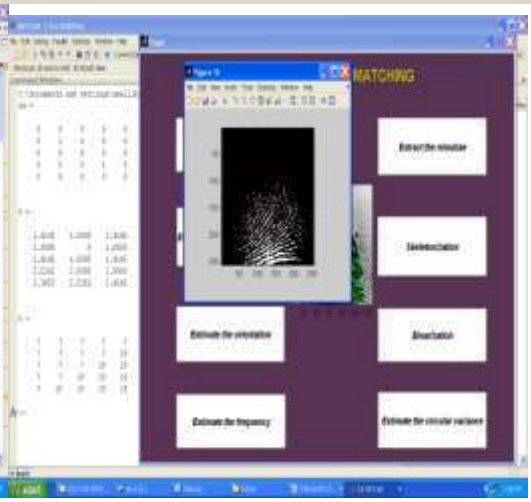


Fig. 13: Estimation of circular variance.

CONCLUSION

This work suggests a prototype which is robust and secure for Fingerprint Matching. This work has two important operations in pre-processing stage as Histogram Equalization, and Selection of ROI. These two operations make this algorithm efficient. The Histogram Equalization enhanced the quality of Input-image, which actually help to produce accurate calculation. This research concludes that the Fingerprint Verification is possible even the quality of the fingerprint image got affected. The experimental results show that the proposed method gives higher matching scores compared to the conventional minutiae based one. Hence we can conclude that proposed ridge features give additional information for fingerprint matching with little increment of template size. We will try to incorporate these features into the state-of-the-art minutiae based matchers for further improvement of the matching performance. Also, our matching method needs to be improved for images with a small foreground as, area and those of low quality. We will develop the global knowledge of the fingerprints such as singular point position, to enhance the matching accuracy.

CONFLICT OF INTEREST
None

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] J S Bart°un˘ek, M Nilsson, J Nordberg, and I Claesson, [2006] Adaptive fingerprint binarization by frequency domain analysis, in Proc. IEEE40th Asilomar Conf. Signals, Syst. Comput. 598–602.

[2] R Cappelli, D Maio, D Maltoni, J Wayman, and A Jain, [2006] Performance evaluation of fingerprint verification systems, IEEE Trans. Pattern Anal.Mach.Intell. 28(01): 3–18.

[3] Y Chen, S Das, and A Jain,[2005] Fingerprint quality indices for predicting authentication performance, in Proc. 5th Int. Conf. Audio-Video-Based Biometric Person Authent. 160–170.

[4] AntoBennet, M, Sankar Babu G, Natarajan S, [2015] Reverse Room Techniques for Irreversible Data Hiding, Journal of Chemical and Pharmaceutical Sciences 08(03): 469-475.

[5] AntoBennet, M Sankaranarayanan S, Sankar Babu G, [2015] Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis, Journal of Chemical and Pharmaceutical Sciences 08(03): 571-576.

[6] AntoBennet, M, Suresh R, Mohamed Sulaiman S, [2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", Journal of Chemical and Pharmaceutical Research 07(08): 291-299.

[7] AntoBennet, M & JacobRaglend, [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, Journal of Computer Science, 8(09): 1447-1454.

[8] AntoBennet, M & JacobRaglend, [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering, European Journal of Scientific Research, 66(01):120-129.

[9 AntoBennet, M , Resmi R Nair, Mahalakshmi V, Janakiraman G [2016] Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography", Indian Journal of Science and Technology,9 (02):01-07.

[10] AntoBennet, M & JacobRaglend, [2012] A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, Journal of Applied Sciences Research,8(05): 2429-2438.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

CLASSIFICATION OF ARTERY AND VEIN BY AUTOMATIC GRAPH GENERATION USING LDA CLASSIFIER

M Anto Bennet[1*], TR Dinesh Kumar [2], G Sankar babu [3], M Pooja [4], K Anusuya[5], AP Kokila[6]

[1]Professor,[2,3]Asst.professor,[4,5,6]UG Students,Department of ECE,VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

The classification of retinal vessels into Artery/Vein (A/V) is an important phase for automating the detection of vascular changes, and for the calculation of characteristic signs associated with several systemic diseases such as diabetes, hypertension, and other cardiovascular conditions. It presents an automatic approach for A/V classification based on the analysis of a graph extracted from the retinal vasculature. The proposed method classifies the entire vascular tree deciding on the type of each intersection point (graph nodes) and assigning one of two labels to each vessel segment (graph links). Final classification of a vessel segment as A/V is performed through the combination of the graph-based labeling results with a set of intensity features. The obtained results are compared with the three different data set like DRIVE dataset, INSPIRE dataset. Retinal vessels are affected by several systemic diseases namely diabetes, hypertension, and vascular disorders. In diabetic retinopathy, the blood vessels often show abnormalities at early stages as well as vessel diameter alterations. Changes in retinal blood vessels, such as significant dilatation and elongation of main arteries, veins, and their branches are also frequently associated with hypertension and other cardiovascular pathologies.

KEY WORDS
Artery/Vein (A/V),Optic Disc (OD),Structure-based SampleConsensus (STRUCT-SAC) algorithm, Retinopathy ofPrematurity (ROP)

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

Automatic detection of the retinopathy in eye fundus images using digital image analysis method has huge potential benefits allowing the examination of a large number of images in less time with lower cost and reduced subjectivity than current observer based techniques. Another advantage is the possibility to perform the automated screening for pathology condition such as diabetic retinopathy in order to reduce the workload required of trained manual graders. Retinal vessels are affected by several systemic diseases, namely diabetes, hypertension, and vascular disorders. In diabetic retinopathy, the blood vessels often show abnormalities at early stages, as well as vessel diameter alterations. Changes in retinal blood vessels, such as significant dilatation and elongation of main arteries, veins, and their branches are also frequently associated with hypertension and other cardiovascular pathologies [1,2,3]. Several characteristic signs associated with vascular changes are measured, aiming at assessing the stage and severity of some retinal conditions. generalized arteriolar narrowing, which is inversely related to higher blood pressure levels is usually expressed by the Arteriolar-to-Venular diameter Ratio (AVR). The Atherosclerosis Risk In Communities (ARIC) study previously showed that a smaller retinal AVR might be an independent predictor of incident stroke in middle aged individuals. The AVR value can also be an indicator of other diseases, like diabetic retinopathy and retinopathy of prematurity. Among other image processing operations, the estimation of AVR requires vessel segmentation, accurate vessel width measurement, and Artery/Vein (A/V) classification. Therefore, any automatic AVR measurement system must accurately identify which vessels are arteries and which are veins, since slight classification errors can have a large influence on the final value. It is estimated that about 10% of the population over the age of 40 are affected with diabetes and about 20% of this group will develop some form of diabetic complications in the eye. With the number rising every year, Singapore is one of the countries with the highest rate of diabetes in the world[4,5,6].Several works on vessel classification have been proposed but automated classification of retinal vessels into arteries and veins has received limited attention, and is still an open task in the retinal image analysis field. In recent years, graphs have emerged as a unified representation for image analysis, and graph-based methods have been used for retinal vessel segmentation, retinal image registration, and retinal vessel classification. The graph extracted from the segmented retinal vasculature is analyzed to decide on the type of intersection points (graph nodes), and afterwards one of two labels is assigned to each vessel segment (graph links). Finally, intensity features of the vessel segments are measured for assigning the final artery/vein class [7].

MATERIALS AND METHODS

The method proposed in this follows a graph-based approach, where mostly focus on a characteristic of the retinal vessel tree that, at least in the region near the optic disc, veins rarely cross veins and arteries rarely cross arteries. Based on this assumption it may define different types of intersection points: bifurcation, crossing, meeting, and connecting points. A bifurcation point is an intersection point where a vessel bifurcates to narrower parts. In a crossing point a vein and an artery cross each other. In a meeting point the two types of vessels meet each other without crossing, while a connecting point connects different parts of the same vessel. The classification of arteries and veins inretinal images is essential for the automated assessment of vascular changes. The decision on the type of the intersection points are made based on the geometrical analysis of the graph representation of the vascular structures. An Automatic Graph Generation Algorithm issued for classified Artery and Veins. This method uses additional

information extracted from a graph which represents the vascular network. This method is able to classify the whole vascular tree and does not restrict the classification to specific regions of interest, normally around the optic disc. The process of artery and vein classification from the retinal vessels for the automatic detection of vascular changes and for the calculation of characteristic signs associated with the several systematic diseases such as diabetics, hypertension and cardio vascular conditions are shown in [Fig. 1]

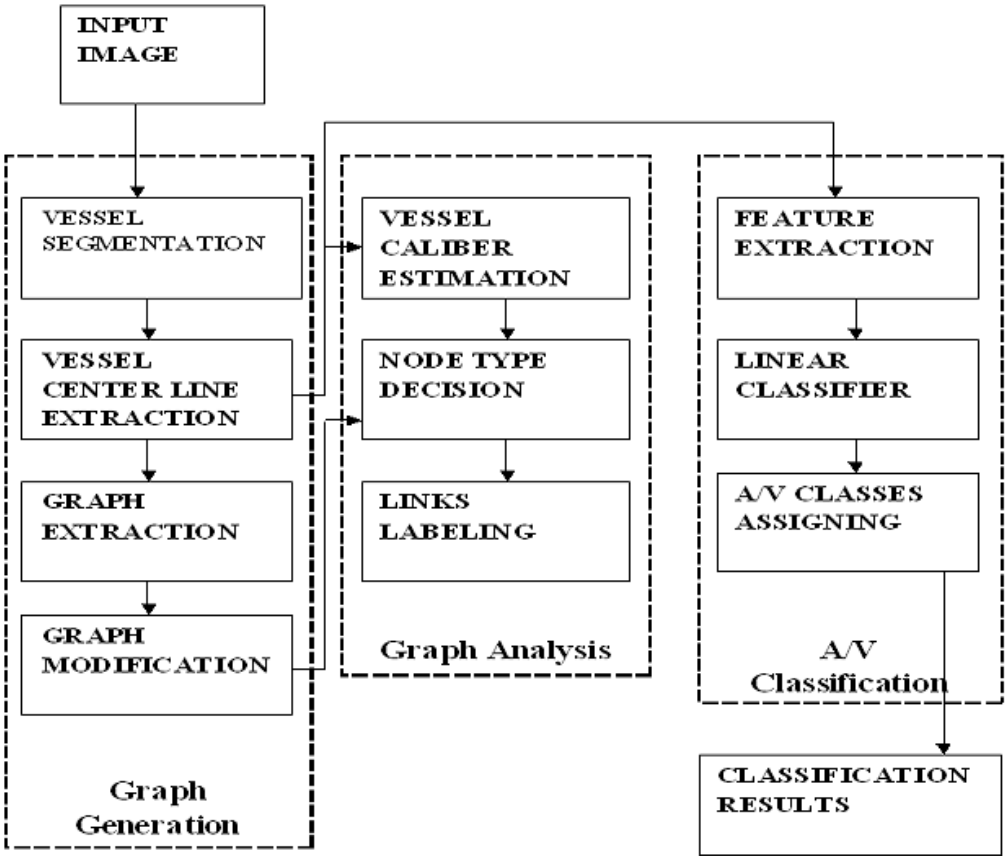


Fig. 1: Block Diagram of the artery and vein classification

Segmentation

The vessel segmentation result is used for extracting the graph and also for estimating vessel calibers. The method proposed was used for segmenting the retinal vasculature, after being adapted for the segmentation of high resolution images. This method follows a pixel processing-based approach with three phases. The first one is the pre-processing phase, where the intensity is normalized by subtracting an estimation of the image background, obtained by filtering with a large arithmetic mean kernel. In the next phase, centerline candidates are detected using information provided from a set of four directional Difference of Offset Gaussian filters, then connected into segments by a region growing process, and finally these segments are validated based on their intensity and length characteristics. The third phase is vessel segmentation, where multiscale morphological vessel enhancement and reconstruction approaches are followed to generate binary maps of the vessels at four scales. The final image with the segmented vessels is obtained by iteratively combining the centerline image with the set of images that resulted from the vessel reconstruction.

Thresholding

The simplest method of image segmentation is called the thresholding method. This method is based on a clip-level (or a threshold value) to turn a gray scale image into a binary image. There is also a balanced histogram thresholding. The key of this method is to select the threshold value (or values when multiple-levels are selected). Several popular methods are used in industry including the maximum entropy method, Otsu's method (maximum variance), and k-means clustering. Recently, methods have been developed for thresholding Computed Tomography (CT) images. The key idea is that, unlike Otsu's method, the thresholds are derived from the radiographs instead of the (reconstructed).

Centerline Extraction

The centerline image is obtained by applying an iterative thinning algorithm described in to the vessel segmentation result. This algorithm removes border pixels until the object shrinks to a minimally connected stroke.

Graph Generation

The graph nodes are extracted from the centerline image by finding the intersection points (pixels with more than two neighbors) and the endpoints or terminal points (pixels with just one neighbor). In order to find the links between nodes (vessel segments), all the intersection points and their neighbors are removed from the centerline image and asresult image with separate components is obtained which are the vessel segments. Next, each vessel segment is represented by a link between two nodes.

Graph Analysis

The output of the graph analysis phase is a decision on the type of the nodes. The links in each subgraph (i) are labeled with one of two distinct labels (Ci1 and Ci2). In this phase it is not yet able to determine whether each label corresponds to an artery class or to a vein class. The A/V classes will be assigned to these subgraphs only in the last classification phase. The node classification algorithm starts by extracting the following node information: the number of links connected to each node (node degree), the orientation of each link, the angles between the links, the vessel caliber at each link, and the degree of adjacent nodes.

A/V Classification

For each centerline pixel, the 30 features listed are measured and normalized to zero mean and unit standard deviation. Some of these features were used previously in, the most commonly used classifiers, namely Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbor (kNN), on the INSPIRE-AVR dataset. For feature selection, we have used sequential forward floating selection, which starts with an empty feature set and adds or removes features when this improves the performance of the classifier.The trained classifier is used for assigning the A/V classes to each one of the sub graph labels. First, each centerline pixel is classified into A or V classes, then for each label (Ci j , j = 1, 2) in sub graph i , the probability of its being an artery is calculated based on the number of associated centerline pixels classified by LDA to be an artery or a vein. The probability of label Ci j to be an artery is Pa(Ci j) = naCi j /(naCi j + nvCi) where naCi j is the number of centerline pixels of a label classified as an artery and nvCi j is the number of centerline pixels classified as a vein. For each pair of labels in each sub graph, the label with higher artery probability will be assigned as an artery class, and the other as a vein class.Finally, to prevent a wrong classification as a result of a wrong graph analysis, to calculate the probability of being an artery or a vein for each link individually. The probability of a link (li) being an artery (Pa (li)) is computed as Pv (li) = nv li / _ na li + nv li _ , and the probability of being a vein ((Pv (li)) is computed as Pv (li) = nv li / _ na li + nv li _ , here na li is the number of centerline pixels of link (li) classified as an artery and nv li is the number of centerline pixels classified as a vein. If the probability of being an artery is higher than 0.9 (Pa (li)≥ 0.9) then the link will be assigned as an artery, and if Pv (li) ≥ 0.9 then it will be assigned as a vein, without considering the result of the graph analysis.

Feature Extraction

To analyze retinal images described above and extract information, features need to be mined. These features include blood vessels, microaneurysms and the optic disc.

Blood Vessel Features

Lines are composed of edges. Awcock and Thomas defined an edge in a digitized image as a sequence of connected edge points where an edge is characterized by abrupt changes in intensity indicating the boundary between two regions in an image. Based on this, a line according to their definition is a region of constant intensity found between two edges which act as a boundary for the line. Blood vessels in the retina match the criteria of a line shown in fig 2. It shows two examples of blood vessels. Throughout the retina the major blood vessels supply the capillaries that run into the neural tissue. Capillaries are found running through all parts of the retina from the nerve fiber layer to the outer layer. There are two sources of blood supply to the mammalian retina: the central retinal artery and the choroidal blood vessels. The choroid receives the greatest blood flow (65%-85%) and is vital for the maintenance of the outer retina (particularly the photoreceptors) and the remaining 20%-30% flows to the retina through the central retinal artery from the optic nerve head to nourish the inner retinal layers .

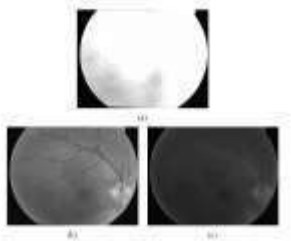


Fig. 2: The corresponding color bands Red (a), Green (b) and Blue (c) of the color retinal image

Microaneurysm Features

Microaneurysms are the dilation of retinal capillaries. They are round intra-retinal lesions ranging from 10 to 100 micrometers in size and red in color. The cross-section of a microaneurysm exhibits a Gaussian distribution. [Fig.3] illustrates examples of different microaneurysms taken from color retinal images. The top part shows their original format while the bottom depicts them in the green channel (so their shape is more visible). Researchers at the European Association for the Study of Diabetes 45th Annual Meeting in Vienna, Austria, reported that an increase in the number of retinal microaneurysms is associated with worse retinopathy prognosis in patients with Type 1 or 2 diabetes.

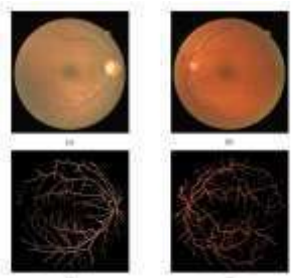


Fig.3: Fundus images (a), (b) (c), (d) and its corresponding vessel map

Optic Disc Features

The Optic Disc (OD) or optic nerve head, another commonly used name, is a vertical oval with average dimensions of 1.76mm (horizontally) × 1.92mm (vertically), and situated 3-4mm to the nasal side of the fovea. There are no receptors in this part of the retina since all of the axons of the ganglion cells exit the retina to form the optic nerve. In fundus imaging the OD is usually brighter than its surrounding area, and is the convergence of the retinal blood vessel network. This can be seen in [Fig. 4] and [Fig.5] which shows four different ODs.

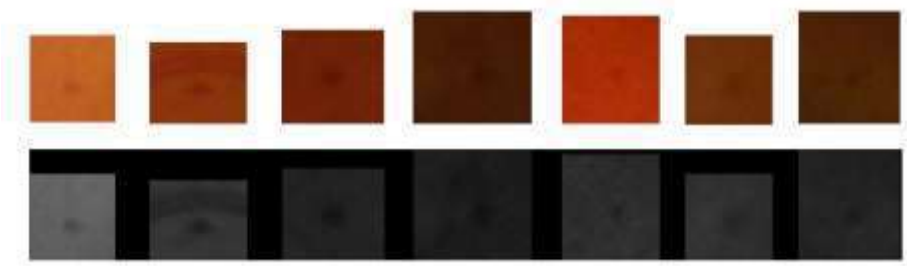


Fig. 4: Examples of different microaneurysms shown on the top with its equivalent green channel on the bottom

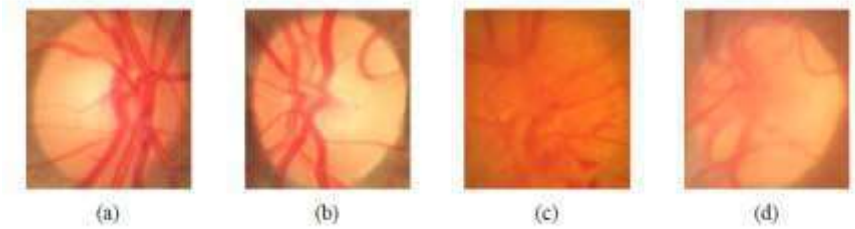


Fig. 5: Four cropped images of the optic disc

Optic Nerve Head Assesment

Assessment of the daaged optic nerve head is both more promising, and superior to measurement or visual field testing for glaucoma screening. Optic nerve head assessment can be done by a trained professional. However, manual assessment is subjective, time consuming and expensive. Therefore, automatic optic nerve head assessment would be very beneficial. Optic nerve-head examination is probably the most important step in the diagnosis of glaucoma and is also extremely important in monitoring patients with established glaucoma. There are several ways to clinically examine the optic nerve head, including direct ophthalmoscopy, indirect ophthalmoscopy, and slit lamp biomicroscopy with contact lenses (such as a Goldman lens), handheld lenses (such as a 78 or 90-diopter lens) or the Hruby lens. Clinical examination of the optic nerve should be performed with similar methodology each and every time it is executed, in order not to miss important aspects of the examination. In my view, examination of the optic nerve head should start with an evaluation of optic disc size, since disc size is extremely important in the interpretation of other optic nerve findings.

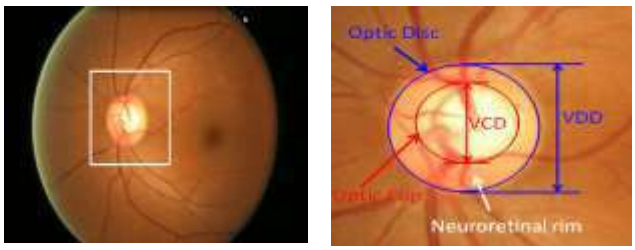


Fig. 6: Major structures of the optic disc

[Fig. 6] shows the optic nerve head or the optic disc (in short, disc) is the location where ganglion cell axons exit the eye to form the optic nerve, through which visual information of the photoreceptors is transmitted to the brain. In 2D images, the disc can be divided into two distinct zones; namely, a central bright zone called the optic cup (in short, cup) and a peripheral region called the neuroretinal rim.

RESULTS

Input image and the background normalised image

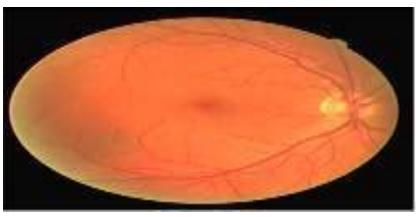


Fig.7: Input image

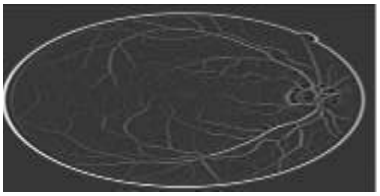


Fig.8: Background normalized image ofretina

The [Fig. 7] is the input image which is given as the input. The pixel size of the input image is 225x224. [Fig. 8] is the background normalized image of the retina. From the input image the background is only subtracted to obtain the background normalized image in order to obtain the clear intensity features.

Fig. 9: Vessel segmentation image of retina

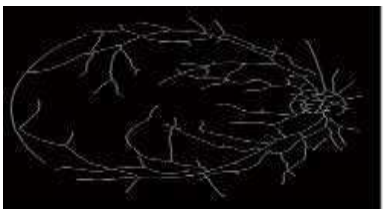


Fig. 10: Centerline extraction image of Retina.

The [Fig.7] is the vessel segmented image of the retina. From the [Fig 5.2] only the vessels are segmented and shown in the [Fig. 8].The [Fig. 9] is the centerline extraction image of the retina.In the [Fig.10] the iterative thinning algorithm is applied and the border pixels are only removed then the vessels are thinned.

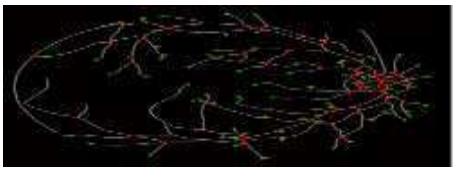


Fig. 11: Intersection points and the

Fig. 12: Graph generationimage of the end point of the retina

The [Fig. 11] is the intersection points and the end point of the retina. In this the intersection point and the end point are plotted. The intersection section point is denoted by red colour and the end point is denoted by green colour. The [Fig. 12] is the graph generation image.In this the graph nodes are extracted from the centerline images by finding the intersection point and the end point.The curved lines are changed into the straight lines in the graph generation image of the retina. Then the intersection point and the links are removed from the centerline image.

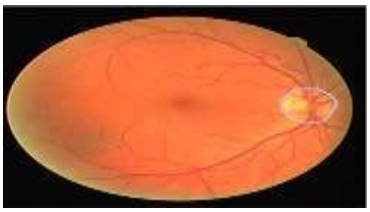


Fig. 13: Optic disk removed image

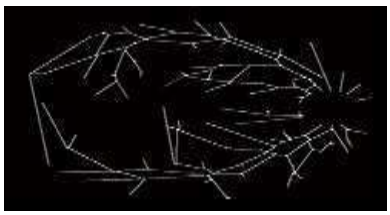


Fig. 14: Graph generation of the retina

The [Fig. 13] is the optic disk removed image. In this figure the optic disk is removed from the center of the image. Then the intersection point and the end point are removed from the centerline image and the graph is obtained as in the [Fig. 14].

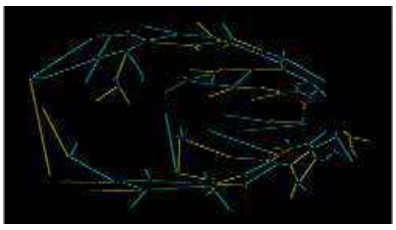


Fig. 15: Graph analysis image

Fig. 16: Classification result

The [Fig. 15] is the graph analysis image. In this [Fig.16] the link between the intersection and the end point are removed and the graph is analysed. Then the obtained graph is classified. The optic disc area usually contains many vessels, but they are not suitable for the artery and vein classification.

Fig.17: Existing method images

Fig. 18: Proposed method image

In the [Fig. 17] the output of the artery and vein classification as in the existing method is obtained. In the [Fig. 18] the proposed method output for the artery and vein classification is obtained with high accuracy when compared with the existing system. In this [Fig. 18] red colour represents the vein and the blue colour represents the artery.

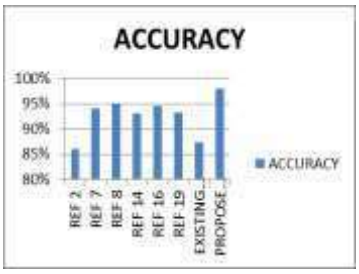


Fig.19: Comparison of accuracy With previous method

Fig.20: Comparison of specificity with previous method

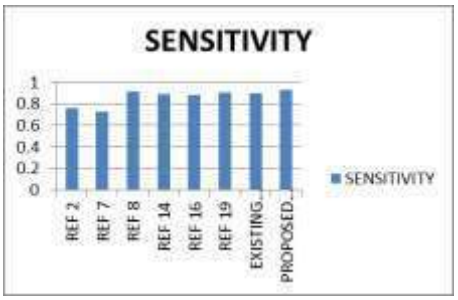


Fig. 21: Comparison of the sensitivity with previous methods

The comparison of accuracy, specificity, sensitivityvalues are compared with the previous methodthat is plotted in the form of graph as shown in [Fig. 19-21].

Table 1: Manual labeling of parameters for artery and vein classification DRIVE dataset

IMAGE	SENSITIVITY	SPECIFICITY	ACCURACY
1	0.797	0.972	0.94
2	0.824	0.971	0.949
3	0.734	0.974	0.941
4	0.783	0.974	0.949
5	0.738	0.980	0.947
6	0.754	0.967	0.936
7	0.686	0.985	0.945

8	0.660	0.983	0.943
9	0.769	0.970	0.946
10	0.717	0.979	0.948
11	0.759	0.975	0.947
12	0.771	0.977	0.951
13	0.804	0.962	0.939
14	0.771	0.979	0.955
15	0.800	0.972	0.954
16	0.779	0.975	0.950
17	0.733	0.980	0.949
18	0.858	0.961	0.949
19	0.906	0.960	0.954
20	0.870	0.954	0.945

By using the drive data sets the accuracy, sensitivity, specificity are computed in drive data set 40 images are used. In that 20 images are used for testing and remaining 20 images are used for training. Manual labeling is performed by the experts first for calculating the sensitivity, accuracy and specificity values are shown in the table. In the proposed method sensitivity, accuracy and specificity values are simulated using automatic graph generation and the results are compared with manual labeling as shown in the [Table 1&2].

Table 2: Proposed classification of artery and vein by labeling the parameters

IMAGE	SENSITIVITY	SPECIFICITY	ACCURACYIN %
1	0.825	0.997	98.1
2	0.814	0.995	97.8
3	0.810	0.98	97.5
4	0.813	0.996	97.0
5	0.811	0.998	97.6
6	0.81	0.996	97.4
7	0.813	0.996	97.7
8	0.815	0.996	97.9
9	0.825	0.997	97.8
10	0.81	0.995	98.2
11	0.815	0.997	98.1
12	0.82	0.998	98.2
13	0.825	0.998	98.8
14	0.812	0.997	98.4
15	0.82	0.99	98.0
16	0.813	0.987	98.1
17	0.81	0.991	98.4
18	0.82	0.994	99.2
19	0.82	0.99	97.9
2020	0.821	0.992	98.1

CONCLUSION

The proposed A/V classification method on the images of three different databases demonstrate the independence of this method in A/V classification of retinal images with different properties, such as differences in size, quality, and camera angle. On the other hand, the high accuracy achieved by our method, especially for the largest arteries and veins, confirm that this A/V classification methodology is reliable for the calculation of several characteristic signs associated with vascular alterations. Further research is planned using the graph that represents the vessel tree and the A/V classification method for AVR calculation, as well as identifying other vascular signs, such as vascular bifurcation angles, branching patterns, and fractal-based features, which can have significant impact on the early detection and follow-up of diseases, namely diabetes, hypertension, and cardiovascular diseases.In the future it will be applied

to feature with edge based artery and vein classification in high accuracy than the proposed system. Further research is planned using the graph that represents the vessel tree and the A/V classification method for AVR calculation, as well as identifying other vascular signs, such as vascular bifurcation angles, branching patterns, and fractal-based features, which can have significant impact on the early sdetection and follow up of diseases, namely diabetes, hypertension, and cardiovascular diseases.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] AntoBennet M, Sankar Babu G, Natarajan S. September [2015] "Reverse Room Techniques for Irreversible Data Hiding", *Journal of Chemical and Pharmaceutical Sciences.* 08(03): 469-475.

[2] AntoBennet M, Sankaranarayanan S, Sankar Babu G. August [2015] "Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis", *Journal of Chemical and Pharmaceutical Sciences* 08(03): 571-576.

[3] AntoBennet M, Suresh R, Mohamed Sulaiman S. August [2015] "Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", *Journal of Chemical and Pharmaceutical Research.* 07(08): 291-299.

[4] AntoBennet M & JacobRaglend. [2012] "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", *Journal of Computer Science*. 8(9):1447-1454.

[5] AntoBennet M & JacobRaglend. [2011] "Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", *European Journal of Scientific Research*. 66(1):120-129.

[6] AntoBennet M , Resmi R, Nair, Mahalakshmi V, Janakiraman G. January [2016] "Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography", *Indian Journal of Science and Technology*. 09 (02):01-07.

[7] AntoBennet M & JacobRaglend. [2012] 'A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach', *Journal of Applied Sciences Research*. 8(5):2429-2438.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

MULTISTAGE FEATURE EXTRACTION OF FINGER VEIN PATTERNS USING GABOR FILTERS

G Sankar Babu[1], N.D.Bobby[2], M Anto Bennet[3], B Shalini[4], K Srilakshmi[5]

[1]Asst.professor[2]Professor,[4,5]UG Students, Department of ECE,VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

[3]Professor,Department of ECEVeltech High Tech Dr.Rangarajan Dr.Sakunthala Engg.College,Avadi,Chennai-600 062, Tamil Nadu, INDIA

ABSTRACT

This paper presents a new approach to improve the performance of finger-vein identification systems presented in the literature. The proposed system simultaneously acquires the finger-vein and low-resolution fingerprint images and combines these two evidences using a novel score-level combination strategy. We examine the previously proposed finger-vein identification approaches and develop a new approach that illustrates it superiority over prior published efforts. The utility of low-resolution fingerprint images acquired from a webcam is examined to ascertain the matching performance from such images. We develop and investigate two new score-level combinations, holistic and nonlinear fusion, and comparatively evaluate them with more popular score-level fusion approaches to ascertain their effectiveness in the proposed system.

KEY WORDS
Automated Fingerprint Identification Systems (AFIS), Region of Interest (ROI)

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

We propose a method of personal identification based on finger-vein patterns. An image of a finger captured under infrared light contains not only the vein pattern but also irregular shading produced by the various thicknesses of the finger bones and muscles. Radon transforms and neural networks are used for the identification of finger vein patterns [1-3]. The proposed method extracts the finger-vein pattern from the unclear image by using line tracking that starts from various positions. Experimental results show that it achieves robust pattern extraction, and the equal error rate was 0.145% in personal identification.In this work is on the development of new approaches for both the finger-vein and finger texture identification, which achieves significantly improved performance over previously proposed approaches the unconstrained finger texture imaging with a low-resolution webcam presents high rotational and translational variations [3-6]. A robust image normalization scheme is developed; rotational and translational variations are also accommodated in our matching strategy, which results in significantly improved performance. This paper investigates two new score-level combination approaches, holistic and nonlinear fusion, for combining finger-vein and finger texture matching scores.A Gabor filter is a linear filter whose impulse response is defined by a harmonic function multiplied by a Gaussian function. Because of the multiplication-convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. Gabor filters are directly related to Gabor wavelets, since they can be designed for number of dilations and rotations. However, in general, expansion is not applied for Gabor wavelets, since this requires computation of biorthogonal wavelets, which may be very time-consuming. Therefore, usually, a filter bank consisting of Gabor filters with various scales and rotations is created [7,8].The Gabor Filters have received considerable attention because the characteristics of certain cells in the visual cortex of some mammals can be approximated by these filters. In addition these filters have been shown to posses optimal localization properties in both spatial and frequency domain and thus are well suited for texture segmentation problems. Gabor filters have been used in many applications, such as texture segmentation, target detection, fractal dimension management, document analysis, edge detection, retina identification, and image coding and image representation. A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope.

$$G(x,y) = s(x,y)\, g(x,y) \quad (1)$$

where s(x,y) is complex sinusoid and g(x,y) is 2D gaussian envelope

$$s(x,y) = \exp . \quad (2)$$

$$g(x,y) = \exp \quad (3)$$

and characterize the spatial extent and bandwidth of along the respective axes, and are the shifting frequency parameters in the frequency domain.

MATERIALS AND METHODS

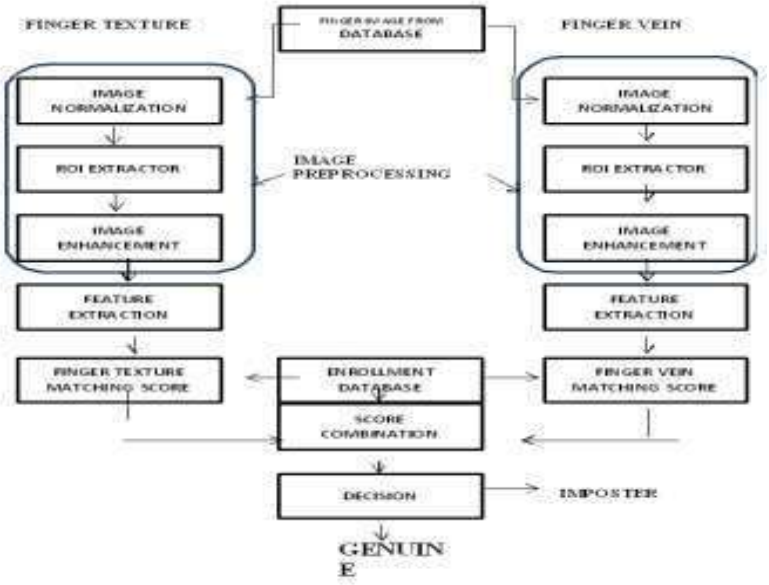


Fig.1: Block Diagram of finger vein & finger texture identification

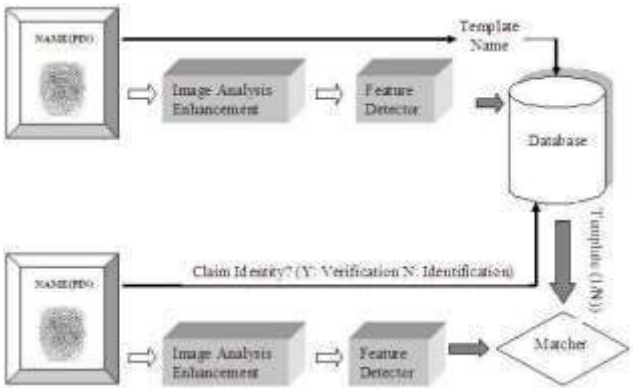


Fig. 2: Data Flow Diagram

Module 1: Finger Vein Identification

Image Normalization

Normalization is a process that changes the range of pixel intensity values. In this, the image is subjected to binarization with threshold value of 230. Sobel edge detector is applied to the image to the remove background portions connected to it. Eliminating the number of connected white pixels being less than a threshold, to obtain the binary mask. Binarization is a method of transforming grayscale image pixels into either black or white pixels by selecting a threshold. The process can be fulfilled using a multitude of techniques. Binarization is relatively easy to achieve compared with other image processing techniques.Finger print Image Binarization is to transform the 8-bit Gray fingerprint image to a 1-bit image with 0-value for ridges and 1-value for furrows. After the operation, ridges in the fingerprint are highlighted with black color while furrows are white. A locally adaptive binarization method is performed to binarize the fingerprint image shown in [Fig. 1].

ROI Extractor

In the finger images, there are many unwanted regions (that cannot be taken for analysis) has been removed by choosing the interested area in that image. The useful area is said to be “Region of Interest”. The obtained binary mask is used to segment the ROI (Region of Interest) from the original finger-vein

image. The orientation of the image is determined to remove the low quality images that present in finger vein image. This orientation is used for the rotational alignment of the ROI in vein image.

Fingerprint Image Segmentation

In general, only a Region of Interest (ROI) is useful to be recognized for each fingerprint image. The image area without effective ridges and furrows is first discarded since it only holds background information. Then the bound of the remaining effective area is sketched out since the minutia in the bound region is confusing with that spurious minutia that is generated when the ridges are out of the sensor. To extract the ROI, a two-step method is used. The first step is block direction estimation and direction variety check, while the second is intrigued from some Morphological methods.

Block direction estimation

The direction for each block of the fingerprint image with W x W in size(W is 16 pixels by default)is estimated. The algorithm is:

I.The gradient values along x-direction (gx) and y-direction (gy) for each pixel of the block is calculated. Two Sobel filters are used to fulfill the task.

II. For each block, following formula is used to get the Least Square approximation of the block direction.

$$tg2\beta = 2\sum\sum(g_x * g_y)/\sum\sum(g_x^2 - g_y^2) \quad (7)$$

The formula is easy to understand by regarding gradient values along x-direction and y-direction as cosine value and sine value. So the tangent value of the block direction is estimated nearly the same as the way illustrated by the following formula.

$$tg2\theta = 2\sin\theta\cos\theta/(\cos^2\theta - \sin^2\theta) \quad (8)$$

After the estimation of each block direction, those blocks without significant information on ridges and furrows are discarded based on the following formulas:

$$E = \{2\sum\sum(g_x * g_y) + \sum\sum(g_x^2 - g_y^2)\}/W * W * \sum\sum(g_x^2 + g_y^2) \quad (9)$$

For each block, if its certainty level E is below a threshold, then the block is regarded as a background block.

ROI extraction by Morphological operations

Two Morphological operations called 'OPEN' and 'CLOSE' are adopted. The 'OPEN' operation can expand images and remove peaks introduced by background noise. The 'CLOSE' operation can shrink images and eliminate small cavities. The bound is the subtraction of the closed area from the opened area. Then the algorithm throws away those leftmost, rightmost, uppermost and bottommost blocks out of the bound so as to get the tightly bounded region just containing the bound and inner area.

Image Enhancement

The acquired image is thin and it is not clear. So the image is enhanced by using bicubic interpolation for better visualization. Fingerprint Image enhancement is to make the image clearer for easy further operations. Since the fingerprint images acquired from sensors or other Medias are not assured with perfect quality, those enhancement methods, for increasing thecontrast between ridges and furrows and for connectingthe falsebroken points of ridges due to insufficient amount of ink, are very useful for keep a higher accuracy to fingerprint recognition. The Method adopted in fingerprint recognition system is Histogram Equalization. Histogram equalization is to expand the pixel value distribution of an image so as to increase the perceptional information. The original histogram of a fingerprint image has the bimodal type. The histogram after the histogram equalization occupies all the range from 0 to 255 and the visualization effect is enhanced.

Module 2: Finger Texture Identification

Localization and Normalization

In texture preprocessing, Sobel edge detector is used to obtain the edge map and localize the finger boundaries. This edge map is isolated with noise and it can be removed from the area threshold. Such noise is eliminated from the area thresholding, i.e., if the number of consecutive connected pixels is less than the threshold. The slope of the upper finger boundary is then estimated. This slope is used to automatically localize a fixed rectangular area, which begins at a distance of 20 pixels from the upper

finger boundary and is aligned along its estimated slope. We extract a fixed 400 160 pixel area, at a distance of 85 and 50 pixels, respectively, from the lower and right boundaries, from this rectangular region. This 400 160 pixel image is then used as the finger texture image for the identification.

Image Enhancement

In Image enhancement, finger texture image is subjected to median filtering to eliminate the impulsive noise. The resulting images have low contrast and uneven illumination. Therefore obtain the background illumination image from the average of pixels in 10 10 pixel image subblocks and bicubic interpolation.The resulting image is subtracted from the median-filtered finger texture image and then subjected to histogram equalization.

Finger Vein and Texture Image Feature Extraction

Gabor filter is used for finger vein and texture image feature extraction. Gabor filters optimally capture both local orientation and frequency information from a fingerprint image. Bytuning a Gabor filter to specific frequency and direction, the local frequency and orientation information can beobtained. We have creating the Gabor with specified orientations and these Gabor filter is convolved with the enhanced image to remove the unwanted regions other than the vein and texture regions. In vein images, the extracted vein images are further processed into morphological top-hat operation for obtaining the clear vein patterns shown in [Fig. 2].

Module 3: Finger Texture Identification

Finger Vein and Texture Matching

In that, the matcher block predicts that the vein and texture image is matched with the database. The database contains the features of all vein and texture images.
For matching, two steps has been done

- Extract features
- Match features

These two steps are done by using Matlab in built commands. Vein regions extracted from the image are stored in database.

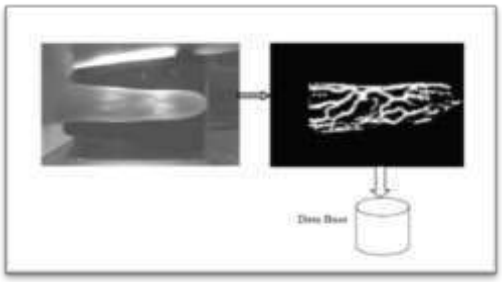


Fig. 3: Data base

Vein Matching

The features extracted from finger vein images are already stored in a database. The features of the input image are matched with all the extracted veins in the database to check whether the input image is matched with any one of the extracted veins.

- If the input image is matched with any one of the extracted veins, the message box will be opened and display “vein matched”.
- If the input image is not matched with any one of the extracted veins, the message box will be opened and display “vein not matched”.

Texture Matching

The features extracted from finger texture images are stored in the same database. The features of the input image are matched with all the extracted texture in the database to check whether the input image is matched with any one of the extracted textures.

- If the input image is matched with any one of the extracted textures, the message box will be opened and display “texture matched”.

- If the input image is not matched with any one of the extracted textures, the message box will be opened and display “texture not matched”.

Module 4: Score Combination

In score level combination, two techniques are used.

- Holistic fusion
- Nonlinear fusion

These two techniques are used to combine the resultant finger vein and texture images. The result of this fusion is used to check whether the fingerprint is genuine or not.

Holistic Fusion

This approach is developed and investigated to utilize the prior knowledge in the dynamic combination of matching scores. Let , and represent the matching score from finger vein, finger texture, and combined score, respectively, and this holistic rule of score combination is given below: (10)The above equation can also be written as,(11).

By using this equation, the final combined scores have a similar trend as the score from vein matching, i.e., when the score from finger-vein matching is high, the fused score will also become high and vice versa. Factor is selected to reflect the reliability of each modality or matching score. We choose the matching score from finger vein as the controlling factor since the performance of finger-vein matching is more stable, as compared with that of the texture.

Nonlinear Fusion

This nonlinear score combination attempts to dynamically adjust the combined score according to the degrees of consistency between the two matching scores and is illustrated as below: (12)

Where is a positive constant and is fixed to 1 in our experiments and is selected in the range of [1,2].

RESULTS

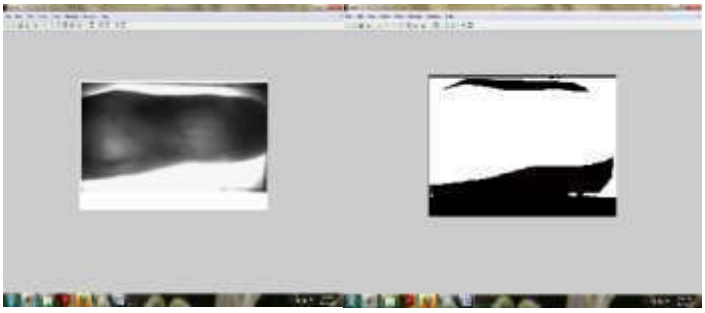


Fig. 4: Original image

Fig. 5: Binarised image

Fig. 6: Freehand selection for ROI Extraction

Fig. 7: ROI

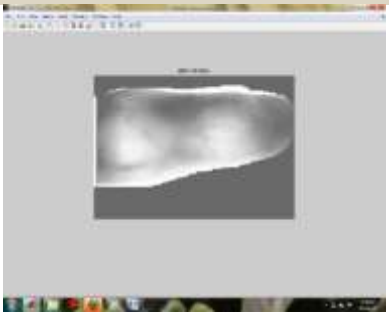


Fig. 8: Enhanced image

Fig. 9: Feature Extracted Vein

Fig. 9: Vein match

Fig.10: Texture input image

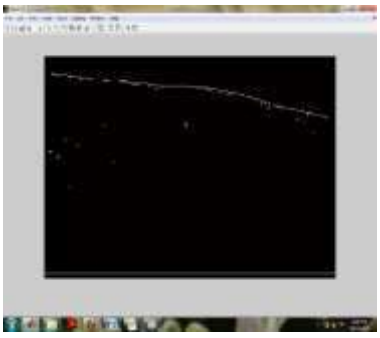


Fig. 11: Edge detection using sobel

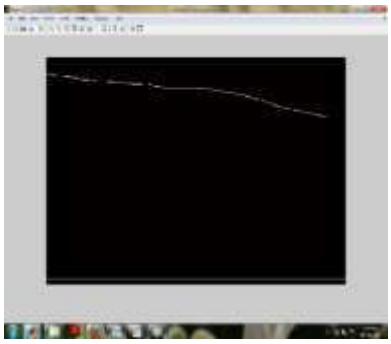


Fig. 12: Image after area thresholding

Fig. 13: Segmented image

Fig. 14: Image median filtered image

Fig. 15: Enhanced Texture image

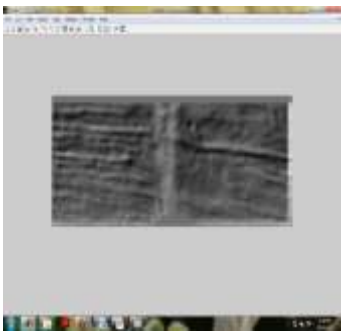


Fig. 16: Feature Extracted image

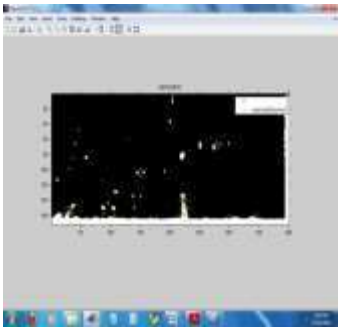


Fig. 17: Texture Match

Fig.18: Matching Scores

In this study ground glass pattern detection of HRCT images is done for better visualization and their analysis. For this we have taken the HRCT images of lungs from Ground glass opacity patients. These are basically RGB images and these are converted to GRAY SCALE images for making further process easy. Then preliminary mask is obtained using Gabor high pass and Gabor low pass filters. Thresholding and Morphological operations such as erosion and dilation are performed to obtain peripheral mask. This peripheral mask contains noise which appears as tiny dots and they are often few pixels wide. For getting accuracy, we are going for post processing technique in which the noise is removed by median filters. This may assist doctors for making decisions for better and quick treatment. The results obtained from the above steps are shown in the following [Fig. 4–18]

CONCLUSION

We have presented a complete and fully automated finger image matching framework by simultaneously utilizing the finger surface and finger subsurface features. We presented a new algorithm for the finger-vein identification, which can more reliably extract the finger vein shape features and achieve much higher accuracy than previously proposed finger-vein identification approaches. Our finger-vein matching scheme works more effectively in more realistic scenarios and leads to a more accurate performance, as demonstrated from the experimental results.In proposed and investigated two new score-level combination approaches, i.e., nonlinear and holistic, for effectively combining simultaneously generated finger-vein and finger texture matching scores. The nonlinear approach consistently performed better than other promising approaches, average, product, weighted sum, Dempster–Shafer, and likelihood-ratio approaches. We examined a complete and fully automated approach for the identification of low resolution finger surface texture images for the performance improvement. This investigation and they obtained results are significant as they point toward the utility of touch less images acquired from the webcam for personal identification and its extension for other utilities such as mobile phones, surveillance cameras, and laptops. Finally, the availability of the acquired database from this paper for the benchmarking/comparison will help further the research efforts in this area. Currently, there is no publicly available database for the performance comparison and research efforts on finger-vein identification.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] AntoBennet M, Sankar Babu G, Natarajan S. September [2015] Reverse Room Techniques for Irreversible Data Hiding, *Journal of Chemical and Pharmaceutical Sciences*. 08(03): 469-475.

[2] AntoBennet M, Sankaranarayanan S, Sankar Babu G. August [2015] Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis, *Journal of Chemical and Pharmaceutical Sciences.* 08(03): 571-576.

[3] AntoBennet M, Suresh R, Mohamed Sulaiman S. August [2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface, *Journal of Chemical and Pharmaceutical Research.* 07(08): 291-299.

[4] AntoBennet M & JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, *Journal of Computer Science*. 8(9): 1447-1454.

[5] AntoBennet M & JacobRaglend. [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", *European Journal of Scientific Research*. 66(1):120-129.

[6] AntoBennet M, Resmi R, Nair, Mahalakshmi V, Janakiraman G. [2016] Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography, *Indian Journal of Science and Technology*. 09 (02):01-07.

[7] AntoBennet M & JacobRaglend. [2012] A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach, *Journal of Applied Sciences Research*. 8(5):2429-2438

[8] Miura N, Nagasaka A, and Miyatake T. [2004] Feature extraction of finger vein patterns based on repeated line tracking and its application to personal identification.

[9] Miura N, Nagasaka A, and Miyatake T. [2005] Extraction of finger-vein patterns using maximum curvature points in image profiles.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

PERFORMANCE ANALYSIS OF AUTOMATIC CLASSIFICATION OF RETINAL VESSELS INTO ARTERIES AND VEINS

M Anto Bennet[1*], S Sankaranarayanan[2], G Sankarbabu[3], A Abinayaa[4], U Mageshwari[5*]

[1]Professor,[2,3]Asst. Professor, [4,5]UG Students,Department of ECE,VEL TECH , Avadi, Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

Our vision reduced in eye due to the presence of Retinal diseases like Exudates (diabetic Retinopathy), Micro aneurysms, and Blood vessel damage. This project mainly concentrates on the symptoms of heart, lung, liver and kidney problems identification using Retinal fundus images. Our proposed work shows that how optic disk elimination and follower the symptom detection. Optic disk is one of the parts which consist of intersection of blood vessels and it also has same characteristics of exudates like yellow color, intensity and contrast. Distinguish the exudates and optic disk is critical one. So only first eliminate the optic disk and follower that exudates detection .This detection method very favorably with existing and promise deployment of these systems. Micro aneurysms are the initial stage of exudates.

KEY WORDS
Confocal Laser Scanning Microscope (CLSM), Fuzzy Inference System (FIS),Adaptive Neuro-fuzzy inference system

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

Infection Blood flow in the microvasculature plays a pivotal role in determining the outcome of injury and repair in inflamed tissue. Real-time observation of the kidney microvasculature, including the glomerular capillary tufts, is extremely difficult because of the methodological limitations of currently available microscope optics. In the present study, we attempted to analyze hemodynamic events that occurred in vivo during microvascular regeneration following destruction of the glomerular capillary tuft, functionally and quantitatively by the use of a real-time confocal laser-scanning microscope (CLSM) system [1]. Blood flow in the microvasculature plays a pivotal role in determining the fate of injury and repair of inflamed tissue. Direct observation of the hemodynamic events occurring in the microcirculation under physiological conditions would allow us to deepen our understanding of the precise mechanisms by which inflammation occurs in the microvasculature. Real-time observation of the kidney microvasculature, including the glomerular capillary tuft, however, is extremely difficult because of the methodological limitations of currently available microscope optics. One strategy to observe the glomerular microvasculature in the rat has been performed using hydronephrotic kidneys. Although pathological changes in the blood vessels caused by this better procedure are relatively minor, especially in rats, hydronephrosis is accompanied by a marked decrease in kidney blood flow. To avoid nonphysiological effects of operative procedures, we introduce an intravital real time confocal laser-scanning microscope (CLSM) system, in combination with fluorescent tracer labeling. We report novel findings during hemodynamic changes in the anti-Thy-1 antibody-induced glomerular aneurysms by surveying with this new equipment. Anti-Thy-1.1 nephritis is a good model for analyzing the hemodynamic changes of glomeruli during the course from destruction of microvasculature to resolution of glomerular architecture [2]. The application of CLSM enables the restitution of glomerular and periglomerular hemodynamics, following mesangial damage to be examined. The human lungs are the organs of respiration in humans. Humans have two lungs, a right lung and a left lung. The right lung consists of three lobes while the left lung is slightly smaller consisting of only two lobes (the left lung has a "cardiac notch" allowing space for the heart within the chest). Together, the lungs contain approximately 2,400 kilometers (1,500 mi) of airways and 300 to 500 million alveoli, having a total surface area of about 70 square meters (750 sq ft) to 100 square metres (1076.39 sq ft) (8,4 x 8,4 m) in adults roughly the same area as one side of a tennis court. Furthermore, if all of the capillaries that surround the alveoli were unwound and laid end to end, they would extend for about 992 kilometres (616 mi). The lungs together weigh approximately 2.3 kilograms, with the right lung weighing more than the left. The pleural cavity is the potential space between the two serous membranes, (pleurae) of the lungs; the parietal pleura, lining the inner wall of the thoracic cage, and the visceral pleura, lining the organs themselves–the lungs. The respiratory system includes the conducting zone, which consists of all parts of the airway that conducts air into the lungs.The parenchyma of the lung, only relates to the functional alveolar tissue, but the term is often used to refer to all lung tissue, including the respiratory bronchioles, alveolar ducts, terminal bronchioles, and all connecting tissues [3]. The trachea divides at a junction the carina of trachea, to give a right bronchus and a left bronchus, and this is usually at the level of the fifth thoracic vertebra. The conducting zone contains the trachea, the bronchi, the bronchioles, and the terminal bronchioles.The respiratory system contains the respiratory bronchioles, the alveolar ducts, and the alveoli. The conducting zone and the respiratory components, except the alveoli, are made up of airways with gas exchange only taking place in the alveoli of the respiratory system. The conducting zone is reinforced with cartilage in order to hold open the airways. Air is warmed to 37 °C (99 °F), humidified and cleansed by the conduction zone; particles from the air being removed by the cilia which are located on the walls of all the passageways[4]. The lungs are surrounded and protected by the rib cage shown in [Fig.1].

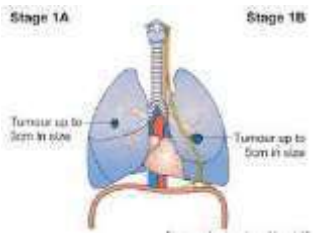


Fig.1: Diagram of Lung

Fig.2: Retina with micro aneurysm formation

MATERIALS AND METHODS

MICROANEURSYM is a tiny area of blood protruding from an artery or vein in the back of eye. These protrusions may open and leak blood into the retinal tissue surrounding it.The problems in Liver are detected through Exudates formation in retina. The liver is a vital organ present in vertebrates and some other animals. It has a wide range of functions, including detoxification, protein synthesis, and production of biochemicals necessary for digestion. The liver is necessary for survival; there is currently no way to compensate for the absence of liver function in the long term, although new liver dialysis techniques can be used in the short term [5].This gland plays a major role in metabolism and has a number of functions in the body, including glycogen storage, decomposition of red blood cells, plasma protein synthesis, hormone production, and detoxification. It lies below the diaphragm in the abdominal-pelvic region of the abdomen. It produces bile, an alkaline compound which aids in digestion via the emulsification of lipids. The liver's highly specialized tissues regulate a wide variety of high-volume biochemical reactions, including the synthesis and breakdown of small and complex molecules, many of which are necessary for normal vital functions shown in [Fig. 2].

Blood flow left lobe liver tumor

The liver supports almost every organ in the body and is vital for survival. Because of its strategic location and multidimensional functions, the liver is also prone to many diseases [6]. The most common include: Infections such as hepatitis A, B, C, D, E, alcohol damage, fatty liver, cirrhosis, cancer, drug damage (particularly by acetaminophen (paracetamol) and cancer drugs).Many diseases of the liver are accompanied by jaundice caused by increased levels of bilirubin in the system[7]. The bilirubin results from the breakup of the hemoglobin of dead red blood cells; normally, the liver removes bilirubin from the blood and excretes it through bile.There are also many pediatric liver diseases including biliary atresia, alpha-1 antitrypsin deficiency, alagille syndrome, progressive familial intrahepatic cholestasis, and Langerhans cell histiocytosis, to name but a few.Diseases that interfere with liver function will lead to derangement of these processes. However, the liver has a great capacity to regenerate and has a large reserve capacity. In most cases, the liver only produces symptoms after extensive damage [8].Liver diseases may be diagnosed by liver function tests, for example, by production of acute phase proteins shown in [Fig. 3].

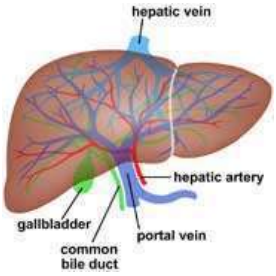


Fig. 3: Diagram of liver

Fig.4: Retina with exudates formation

marked by the masses of white or yellowish layer in the posterior part of the fundus oculi, with deposit of cholestrin and blood debris from retinal hemorrhage shown in [Fig. 4].

Exudates Detection

Exudates are appeared as bright yellow-white deposits on the retina due to the leakage of blood from abnormal vessels.Their shape and size will vary with different diseases according to the stages.The gray scale image is first preprocessed uniformity before the morphological image processing is applied to remove the blood vessels and identify the exudates region.The exudates are detected after removing the border,optical disk and non- exudates area shown in [Fig. 5].

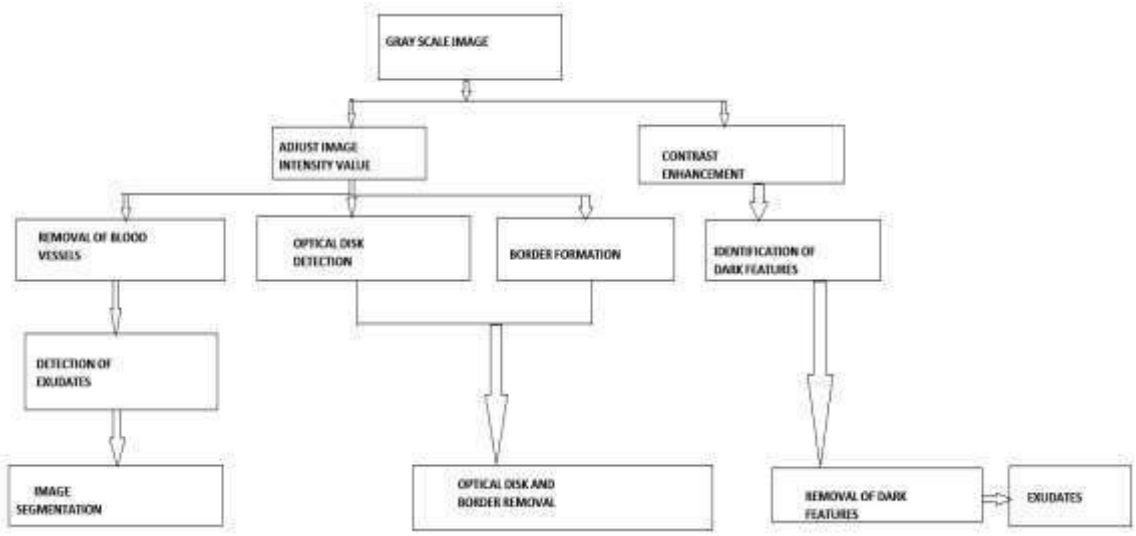


Fig.5: Exudates detection

First the fundus images is first preprocessed to standardize its size 576x720 and the intensity of the gray scale image is then adjusted.Morphological closing which consisted of dilate followed by erode is applied to remove blood vessels. The dilate function expands the exudates area while the erode function removes the blood vessels. The location of the optic disk is detected by the brightest point(s) on the gray scale image.It is usually the maximum value and a circular mask is then created to cover it.The regions of exudates are obtained after the removal of the circular border.Morphological closing is then applied to the image.The dilate function is to fill the exudates while erode function is to expand their sizes.Non-exudates(dark features) are extracted from the gray scale image using function and are represented as binary 1(white) after intensity inversion.AND logic is then applied in the images to detect the exudates.

Experimental Procedure -AND LOGIC

AND logic is used to remove noise for the detection of exudates.Region with exudates are marked out after applying column filter but this includes non-exudates such as hemorrhage and has to be removed as noise.Then removing the non-exudates from the detected regions,the exudates can spots for comparison.These areas (bright features) are represented by binary 0 and the non-exudates (dark features)are rrepresented by binary 1 (white) shown in [Fig. 6] and the corresponding values are shown in [Table 1].

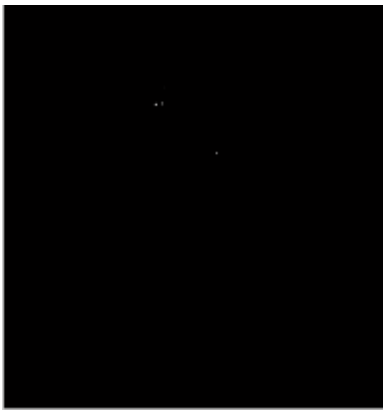


Fig. 6: Exudates

Table 1: Threshold Values For Exudates

AREA	ENERGY	CONTRAST	CORRELATION	HOMOGENITY	OUTPUT
258	0.269534	0.941772713	0.121151789	0.871993035	1
195	0.270413	0.936932181	0.123631065	0.871076244	1
193	0.300596	0.932113324	0.115210245	0.860918304	1
440	0.260851	0.933853832	0.127305377	0.873249401	1
447	0.299253	0.959611786	0.091914368	0.867089627	1
127	0.304575	0.951999122	0.099191051	0.863451239	1
409	0.31617	0.936177946	0.108731471	0.858249425	1
604	0.327915	0.949214444	0.09618857	0.856398872	1
181	0.287229	0.936293525	0.116746738	0.865477467	1
659	0.31674	0.948673083	0.098499257	0.855858107	1
344	0.357489	0.937320011	0.097130299	0.846802964	1
99	0.306869	0.952835068	0.095444089	0.863511905	0
272	0.323525	0.932546543	0.106329258	0.852542795	1
6	0.326471	0.936048735	0.108590078	0.857859424	0
64	0.274644	0.918994111	0.138176178	0.868375894	0
110	0.385856	0.935765869	0.092236365	0.835736322	1
48	0.351474	0.935184607	0.099805383	0.844807925	0

Microaneursym Detection

The gray scale image is first preprocessed uniformity before the morphological image processing is applied to remove the blood vessels and exudates and identify the microaneurysm region.The microaneurysm are detected after removing the border,optical disk and exudates area shown in [Fig. 7].

Fig. 7: Microaneurysm in retina

Experimental Procedure:

First the fundus images are first preprocessed to standardize its size 576x720 and the intensity of the gray scale image is then adjusted. Morphological closing which consisted of dilate followed by erode is applied to remove blood vessels and microaneurysm. The dilate function expands the microaneurysm area while the erode function removes the blood vessels and microaneurysm. The location of the optic disk is detected by the brightest point(s) on the gray scale image.It is usually the maximum value and a circular mask is then created to cover it.The regions of microaneurysm are obtained after the removal of the circular border.Morphological closing is then applied to the image.The dilate function is to fill the microaneurysm while erode function is to expand their sizes.Exudates(dark features) are extracted from the gray scale image using function “im2bw”and are represented as binary 1(white) after intensity inversion.AND logic is then applied in the images to detect the microaneurysm shown in [Fig. 8].

Experimental Procedure -AND LOGIC:

AND logic is used to remove noise for the detection of microaneurysm.Region with microaneurysm are marked out after applying column filter but this includes exudates such as hemorrhage and has to be removed as noise.Then removing the exudates from the detected regions,the exudates can spots for comparison.These areas (bright features) are represented by binary 0 and the non-exudates (dark features)are rrepresented by binary 1 (white) shown in [Fig. 9] and the corresponding values are shown in [Table 2].

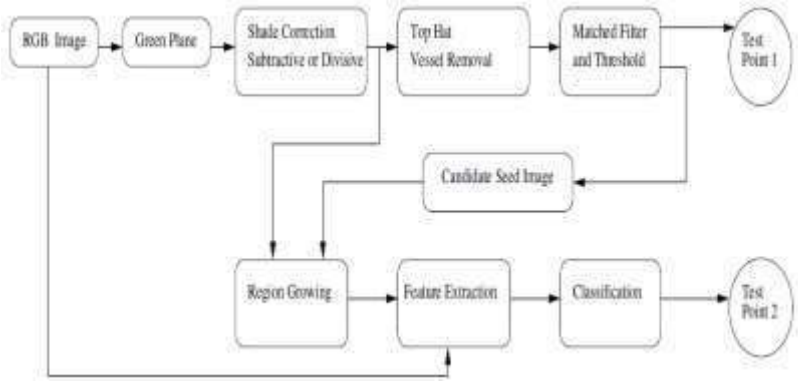


Fig. 8: block diagram for microaneurysm extraction

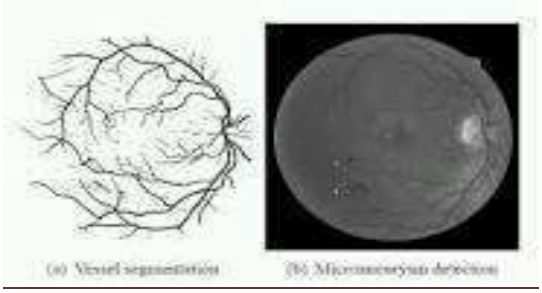


Fig. 9: vessel segmentation and microaneurysm structure.

Table 2: Threshold Values For Microaneurysm

AREA	ENERGY	CONTRAST	CORRELATION	HOMOGENITY	OUTPUT
344	0.357489	0.937320111	0.097130299	0.846802564	1
99	0.306869	0.952835068	0.095444089	0.863511605	0
272	0.321525	0.932546543	0.106329258	0.852542795	1
6	0.326471	0.936048735	0.108590078	0.857859424	0
64	0.274644	0.918994111	0.13817617	0.868375694	0
110	0.385856	0.935765669	0.092236365	0.835736322	1
48	0.351474	0.935184607	0.099805383	0.844807925	0
0	0.273335	0.927134808	0.1276279	0.873611998	0
276	0.362864	0.944921501	0.091267655	0.842319056	1
162	0.265922	0.932088982	0.125012336	0.872845918	1
33	0.274166	0.916474163	0.147605483	0.868018733	0
47	0.279434	0.926004474	0.128673513	0.867605913	0
8	0.340012	0.949237017	0.090816972	0.861026881	0
0	0.348957	0.926848875	0.108737955	0.846607098	0
8	0.323709	0.938626726	0.101828821	0.855786675	0
6	0.329055	0.944470804	0.100862242	0.85348679	0

Training Adaptive Neuro-fuzzy inference system using the ANFIS editor GUI

The acronym ANFIS derives its name from *adaptive neuro-fuzzy inference system*. Using a given input/output data set, the toolbox function anfis constructs a fuzzy inference system (FIS) whose membership function parameters are tuned (adjusted) using either a back propagation algorithm alone or in combination with a least squares type of method. This adjustment allows your fuzzy systems to learn from the data they are modeling shown in [Fig. 10].

The input retinal image is taken and selected for green plane. Retinal vessels are extracted by contrast limited adaptive histogram equalization and morphology and disease is thus observed. This helps clinicians to determine clearly about the type of diseases and thus provide the necessary treatment shown in [Fig. 11].

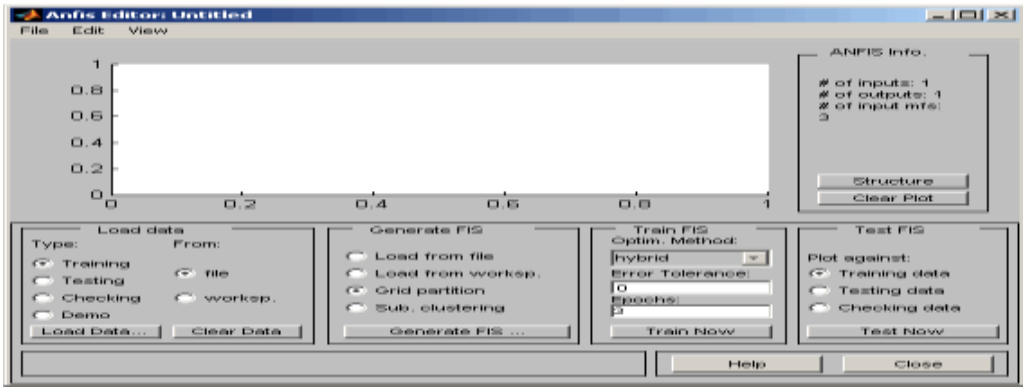


Fig.10: Training Adaptive neuro-fuzzy inference system output

RESULTS

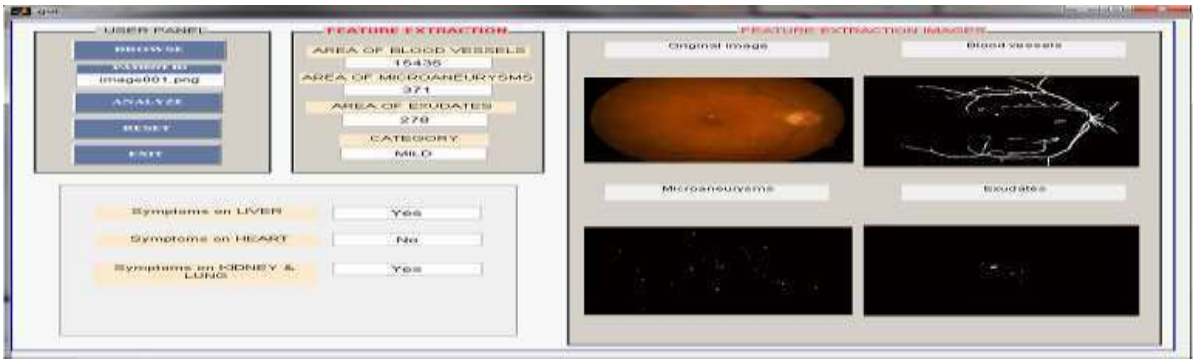


Fig.11: Simulation output

CONCLUSION

Biomedical image processing requires an integrated knowledge in mathematics, statistics,programming and Biology. Different features of the fundus images namely Blood vessels , Exudates and Microaneursyms are extracted using image processing techniques. The values obtained are essential as they represent the image and are necessary in order to classify the images accurately.Based on the result of the classsifier,this project has a sensitivity of 80% and specificity of 20%. It is able to achieve a fairly accurate classification for mild and higher stages, but not for normal class resulting in a possible high false alarm.This might be improved by fine tuning the threshold values used on the images and more images could be used to improve the overall system.In this work, we learnt various techniques of image processing and were able to extract the features namely blood vessels, exxudates and microaneursyms and texture properties like area, energy, contrast, correlation and homogenity from the fundus images.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Vyborny CJ,Giger ML and Nishikawa RM.[2000] "Computer aided detection and diagnosis of breast cancer," Radiologic Clinics Amer N. 38(4):725–740. Jul. 1, 10.1016/S0033-8389(05)70197-4, 0033-8389.

[2] AntoBennet M, Sankar Babu G, Natarajan S. September [2015]. "Reverse Room Techniques for Irreversible Data Hiding", *Journal of Chemical and Pharmaceutical Sciences*. 08(03): 469-475.

[3] AntoBennet M, Sankaranarayanan S, Sankar Babu G. August [2015]. "Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis", *Journal of Chemical and Pharmaceutical Sciences*. 08(03): 571-576.

[4] AntoBennet M, Suresh R, Mohamed Sulaiman S. August [2015] "Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", *Journal of Chemical and Pharmaceutical Research*. 07(08): 291-299.

[5] AntoBennet M.June[2015] "A Novel Effective Refined Histogram For Supervised Texure Classification",

International Journal of Computer & Modern Technology. 01(02):67-73.

[6] AntoBennet M, Srinath R,Raisha Banu A. April [2015] "Development of Deblocking Architectures for block artifact reduction in videos", *International Journal of Applied Engineering Research*. 10(09):6985-6991.

[7] AntoBennet M & JacobRaglend. [2012] "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", *Journal of Computer Science*. 9:1447-1454.

[8] AntoBennet M & JacobRaglend. [2011] "Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", *European Journal of Scientific Research*. 66(1):120-129.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

PERFORMANCE ANALYSIS OF FOREGROUND-ADAPTIVE BACKGROUND SUBTRACTION IN GRAYSCALE VIDEO SEQUENCES

M Anto Bennet[1*], S Lokesh[2], G SankaBabu[3], C Lavanya[4], D Deepa[5], S Srimarthiya[6]

[1]Professor,[2,3]Asst.professor,[4,5,6]UG Students,Department of ECE,VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

The paper proposes efficient motion detection and people counting based on background subtraction using dynamic threshold approach with mathematical morphology. Here these different methods are used effectively for object detection and compare these performance based on accurate detection. Here the techniques frame differences, dynamic threshold based detection will be used. After the object foreground detection, the parameters like speed, velocity motion will be determined. For this, most of previous methods depend on the assumption that the background is static over short time periods. In dynamic threshold based object detection, morphological process and filtering also used effectively for unwanted pixel removal from the background. The background frame will be updated by comparing the current frame intensities with reference frame. Along with this dynamic threshold, mathematical morphology also used which has an ability of greatly attenuating color variations generated by background motions while still highlighting moving objects. Finally the simulated results will be shown that used approximate median with mathematical morphology approach is effective rather than prior background subtraction methods in dynamic texture scenes and performance parameters of moving object such sensitivity, speed and velocity will be evaluated.

KEY WORDS
Fuzzy Color Histogram (FCH), Conventional Color Histogram (CCH), Peak Signal to Noise Ratio(PSNR), Root Mean Square Error(RMSE).

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

There are immediate needs for automated surveillance systems in commercial, law enforcement and military applications. Mounting video cameras is cheap, but finding available human resources to observe the output is expensive. Although surveillance cameras are already prevalent in banks, stores, and parking lots, video data currently is used only "after the fact" as a forensic tool, thus losing its primary benefit as an active, real-time medium. What is needed is continuous 24-hour monitoring of surveillance video to alert security officers to a burglary in progress, or to a suspicious individual loitering in the parking lot, while there is still time to prevent the crime. In addition to the obvious security applications, video surveillance technology has been proposed to measure traffic flow, detect accidents on highways, monitor pedestrian congestion in public spaces, compile consumer demographics in shopping malls and amusement parks, log routine maintenance tasks at nuclear facilities, and count endangered species. The numerous military applications include patrolling national borders, measuring the flow of refugees in troubled areas, monitoring peace treaties, and providing secure perimeters around bases and embassies [1-4].

In 1997, the Defense Advanced Research Projects Agency (DARPA) Information Systems Office began a three-year program to develop Video Surveillance and Monitoring (VSAM) technology. The objective of the VSAM project was to develop automated video understanding technology for use in future urban and battlefield surveillance applications. Technology advances developed under this project enable a single human operator to monitor actives over a broad area using a distributed network of active video sensors [5-7]. The sensor platforms are mainly autonomous, notifying the operator only of salient information as it occurs, and engaging the operator minimally to alter platform operations. A team composed of Carnegie Mellon University Robotics Institute and the Sarnoff Corporation were chosen to lead the technical efforts by developing an end-to-end test bed system demonstrating a wide range of advanced surveillance techniques: real-time moving object detection and tracking from stationary and moving camera platforms, recognition of generic object classes (e.g. human, sedan, truck) and specific object types (e.g. campus police car, FedEx van), object pose estimation with respect to a geospatial site model, active camera control and multi-camera cooperative tracking, human gait analysis, recognition of simple multi-agent activities, real-time data dissemination, data logging and dynamic scene visualization. Twelve other research contracts were awarded to university and industry labs to conduct research in focused technical areas that include human activity recognition, vehicle tracking and counting, airborne surveillance, novel sensor design, and geometric methods for graphical view transfer [8-12].

There are very few studies regarding the wearing and laundering of lab coats in hospitals and medical practice. This study highlights the role of lab coats acting as vector for transmitting health care infections to the patients and the common areas where contamination occurs.

Editor: Dr. K. Sakthisudhan

MATERIALS AND METHODS

Methodology

Proposed System

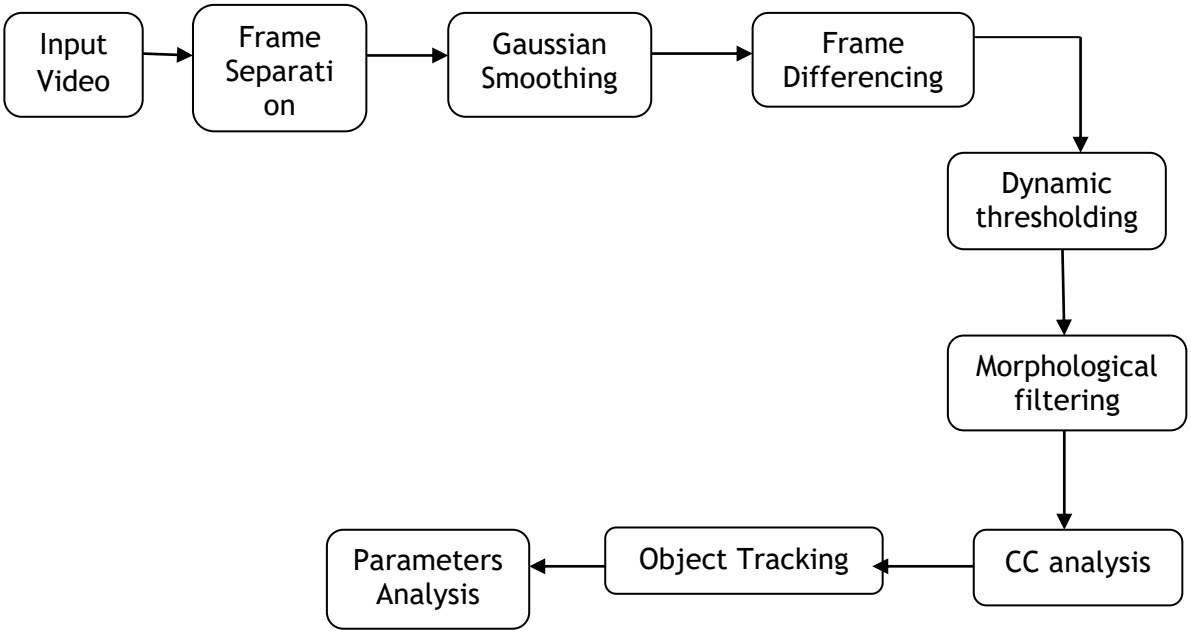


Fig.1: Block Diagram of Entire System

The paper proposes an effective scheme in order to enhance the detection of moving object and their respective tracking. There are various previous methods involved in object detection using background subtraction. The most important phase where they lag is about background motion and shadow elimination. Inspite of these, the system provides better detection and uses simple methodologies. Here the input video is fed to the system. The first process is to separate each frame of the video say for example, A video of 60 frames is converted to a series of 60 images of format .bmp and are stored in a folder. This is done by the Frame separation block. The next step involved in the process is to shape up the frames for further processing which is done by the Gaussian smoothing block. The sizes of the image are corrected to the dimensions that suit right for the process and sent to the Frame differencing block. Here the first image of the video or a series of images are considered to be the background model and each image in the sequence is subtracted from it as a loop of operation. For every loop of subtraction the background model is updated which helps in categorizing the motion involving the shadows and that of the real backgrounds. The next block in the process is the dynamic thresholding. It sets up a threshold value for intensity of the pixels and the size of object. The objects detected are compared with the given thresholds and the objects that do not match the criteria are eliminated from the output. Thus the required foreground is detected in this block. Morphological filtering is the block which involves the addition or removal of pixels in the fore ground detection which would later help in object counts.The CC analysis block compares the object detected with the ground truth (A rough sketch of object to be detected) and helps in measurement of parameters such as sensitivity, Correlation coefficient etc., Then the object tracking block shows the position of detected object in each frame in the video and thus provides an output of effective detection and tracking of object under changing illumination in background and background motions shown in [Fig. 1].

Frame Separation

An Input Video (.avi files) is converted into still images for processing it and to detect the moving objects. These sequences of images gathered from video files by finding the information about it through 'aviinfo' command. These frames are converted into images with help of the command 'frame2im'. Create the name to each images and this process will be continued for all the video frames. The following diagram represents the process flow of this separation shown in [Fig. 2].

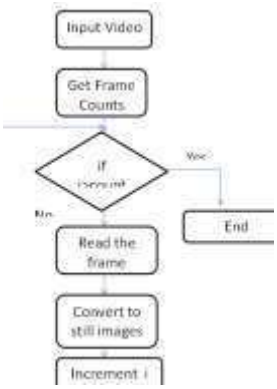


Fig.2: Process flow of Frame separation

Aussian Smoothing

Smoothing of the images can be used to reduce camera noise and remove transient environmental noise such as rain. Many algorithms use a Gaussian blur first to average out fluctuating pixel values to alleviate big differences.Alternatively, when temporal data can be exploited in a video, if a pixel's value is constantly changing over time then it can be assumed it is part of a non-static background object.

Frame Differencing

The moving object will be detected by frame subtraction. The frame subtraction is done by subtracting current frame and previous frame for detecting object from background. Then the background will be updated by comparing the process frame and background frame. This will be continued for all consecutive frames.

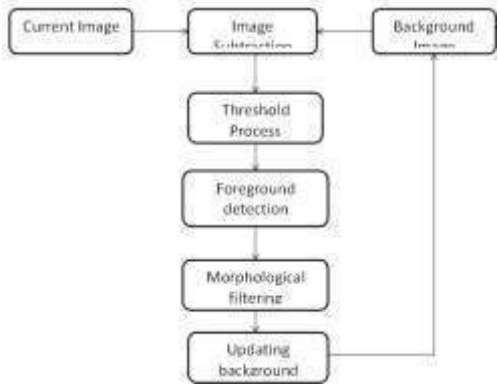


Fig. 3: Process flow of Frame subtraction, threshold process, morphological filtering and background update

Subtraction and Update of the background model

After initialization, temporally subsequent samples are fed to the network. Each incoming pixel P_t of the sequence Frame I_t is compared to the current pixel model C to determine if there exists a weight vector that best matches it. If a best matching weight vector C_m is found, it means that P_t belongs to the background and it is used as the pixel encoding approximation, and the best matching weight vector, together with its neighborhood, is reinforced. Otherwise, if no acceptable matching weight vector exists, we discriminate whether P_t is in the shadow cast by some object or not. In the first case, P_t should be still considered as background, but it should not be used to update the corresponding weight vectors, in order to avoid the reinforcement of shadow information into the background model; in the latter case P_t is detected as belonging to a moving object (foreground) shown in [Fig. 3].

Dynamic Thresholding

The moving object extraction from subtracted frames is done by dynamic thresholding method for foreground detection. The threshold value is set default as an approximate median of objects to be

detected. The current image output is converted into gray scale and cut to desired sizes by the order of rows and columns. Now the difference image obtained is compared with the threshold value given in order to obtain the foreground detection. Each row and column pixel values are related to the assumed threshold and thus the extraction of objects whose pixel values exceed the threshold only are done. The other pixels are eliminated from the outcome.

Morphological Filtering

Morphological techniques probe an image with a small shape or template called a **structuring element**. The structuring element is positioned at all possible locations in the image and it is compared with the corresponding neighbourhood of pixels. Some operations test whether the element "fits" within the neighbourhood, while others test whether it "hits" or intersects the neighbourhood:

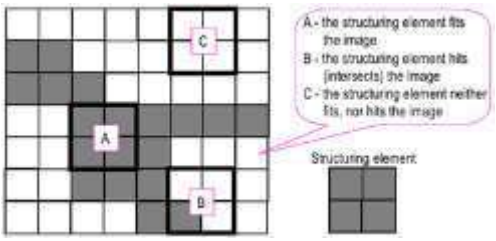


Fig. 4: Probing of an image with a structuring element.

A morphological operation on a binary image creates a new binary image in which the pixel has a non-zero value only if the test is successful at that location in the input image. The structuring element is a small binary image, i.e. a small matrix of pixels, each with a value of zero or one: The matrix dimensions specify the size of the structuring element. The pattern of ones and zeros specifies the shape of the structuring element. An origin of the structuring element is usually one of its pixels, although generally the origin can be outside the structuring element shown in [Fig. 4].

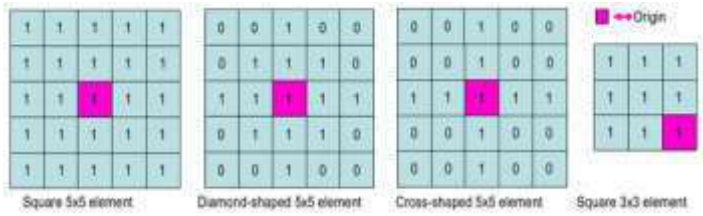
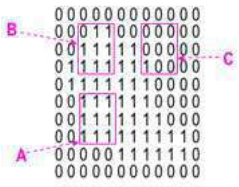


Fig. 5: Examples of simple structuring elements.

A common practice is to have odd dimensions of the structuring matrix and the origin defined as the centre of the matrix. Structuring elements play in morphological image processing the same role as convolution kernels in linear image filtering shown in fig 5.When a structuring element is placed in a binary image, each of its pixels is associated with the corresponding pixel of the neighbourhood under the structuring element. The structuring element is said to fit the image if, for each of its pixels set to 1, the corresponding image pixel is also 1. Similarly, a structuring element is said to hit, or intersect, an image if, at least for one of its pixels set to 1 the corresponding image pixel is also 1. Zero-valued pixels of the structuring element are ignored, i.e. indicate points where the corresponding image value is irrelevant shown in [Fig. 6].



$s_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ $s_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

		A	B	C
fit	s_1	yes	no	no
	s_2	yes	yes	no
hit	s_1	yes	yes	yes
	s_2	yes	yes	no

Fig. 6: Fitting and hitting of binary image with structuring elements s_1 and s_2.

Parameters Evaluation

Velocity

The velocity of object is evaluated based on distance travelled by an object and frame rate

Velocity = Distance travelled / Frame rate (1)

Sensitivity

It measures the proportion of actual positives which are correctly identified in the detection process.

Sensitivity = Tp./(Tp + Fn) (2)

Where, Tp = True Positive: Object pixels correctly classified as object

Fn = False negative: Object pixels incorrectly classified as background.

RESULTS

Input Video

The input video is fed to the system and it is simulated through the matlab software. Here a graphical user interface panel is used in order to display the operations of blocks. The snapshot of the simulated result is shown [Fig. 7].

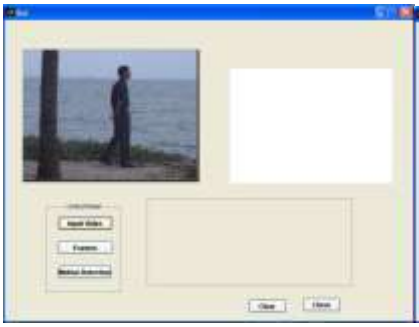


Fig.7: Input video

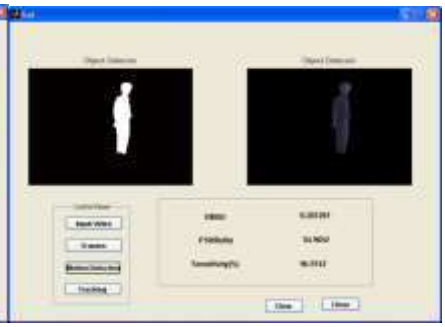


Fig.8: Object detection

Object Detection

The conversion of original frames to grayscale and the detection of the position of moving object in each frame is shown in the snapshot of simulated results. The values of Peak Signal to Noise Ratio (PSNR), Root Mean Square Error(RMSE), Sensitivity of detection are calculated and displayed in the [Fig. 8].

Object Tracking

The tracking of the moving object and the parameters of it such as speed, velocity and object count are determined and displayed in the [Fig.9]. The GUI panel helps in working the module easier. The parameters calculated are helpful in better comparative analysis of the proposed system with the other previously existing systems.

Fig.9: Object tracking

CONCLUSION

The paper presented an efficient motion detection based on background subtraction using frame difference with thresholding and mathematical morphology. It will be enhanced with futures of connected component analysis and morphological filtering for tracking and counting moving objects. After the foreground detection, the parameters like Count, velocity of the motion was estimated and performance of object detection will be measured with sensitivity and correlation using ground truth. Finally the proposed method will be proved that effective for background subtraction in static and dynamic texture scenes compared to prior methods.The system can be programmed in integrated chip and can be inhibited into the camera itself, so that real time detection can be achieved. The threshold value which has been calculated manually in the process based on the approximate median of the object intensity can be automated in future for higher efficiency. More effective threshold techniques can be implemented for low quality videos which would help satellite surveillance applications.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Purohit Kalyan Kumar Hati, Pankaj Kumar Sa, and Banshidhar Majhi. [2013] Intensity Range Based Background Subtraction for Effective Object Detection", IEEE Signal processing letters. 20(8):759-763.

[2] Reddy V, Sanderson C and Lovel Bl. [2013] Improved foreground detection via block-based classifier cascade with probabilistic decision integration," IEEE Trans. Circuits Syst. Video Technol. 23(1): 83–93.

[3] Liu Z, Huang K and Tan T. [2012] Cast shadow removal in a hierarchical manner using MRF, IEEE Trans. Circuits Syst. Video Technol. 22(1):56–66.

[4] Kim W and Kim C. [2012] Background subtraction for dynamic texture scenes using fuzzy color histograms," IEEE Signal Process. Lett. 19 (3):127–130.

[5] Barnich O and Van Droogenbroeck M. [2011] ViBe: A universal background subtraction algorithm for video sequences, IEEE Trans. Image Process. 20(6):1709–1724.

[6] AntoBennet M, Sankar Babu G, Natarajan S. [2015] Reverse Room Techniques for Irreversible Data Hiding", *Journal of Chemical and Pharmaceutical Sciences*. 08(03): 469-475.

[7] AntoBennet M, Sankaranarayanan S, Sankar Babu G. [2015] Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis", *Journal of Chemical and Pharmaceutical Sciences* 08(03): 571-576.

[8] AntoBennet M, Suresh R, Mohamed Sulaiman S. [2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface, *Journal of Chemical and Pharmaceutical Research*. 07(08): 291-299.

[9] AntoBennet M. [2015] A Novel Effective Refined Histogram For Supervised Texure Classification", *International Journal of Computer & Modern Technology*. 01(02):67-73.

[10] AntoBennet M, Srinath R,Raisha Banu A. [2015] "Development of Deblocking Architectures for block artifact reduction in videos", *International Journal of Applied Engineering Research*. 10(09):6985-6991.

[11] AntoBennet M & JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", *Journal of Computer Science*. 8(9):1447-1454.

[12] AntoBennet M & JacobRaglend. [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering" *European Journal of Scientific Research*. 66(1):120-129.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

VLC USING SINGLE INPUT SINGLE OUTPUT STRATEGIES AND MULTIPLE INPUT MULTIPLE OUTPUT ANTENNA STRATEGIES THROUGH OSDM

M Anto Bennet[1*], G Vijayalakshmi[2], M Vijayalakshmi[3], P Shenbagavalli[4], S Saranya[5]

[1]Professor,[2]Asst. Professor,[3,4,5]UG Students,Department of ECE,VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

With increasing demands for faster and more secure wireless communications, there is a pressing need for a new medium of wireless communication as the radio spectrum is already crowded. Visible light is a medium that can address both of these needs. It is a relatively new technology with great potential. This work was completed to develop a working visible light communication system and demonstrate the transmission capabilities of such a system. This year's team set a goal to surpass the previous year's team in transmission speed, range, and size. Of these goals, transmission speed and range were both achieved, while the transmission of a large audio file was deemed not possible based off the difficulties the team encountered while processing large amounts of data.

KEY WORDS
forward voltage tracking (VFT), Pseudo-random binary sequences (PRBS),MIMO(multiple input multiple output).

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

Visible light communication is a viable technology to accommodate the need for faster and better wireless communications in the coming years. The basic idea, is that instead of using traditional methods of communication over cables or radio frequencies, VLC systems send data by turning light on (logic 1) and off (logic 0). This report describes and evaluates the visible light communication system design the team created. As visible light communication technology is relatively new, the team worked on creating a prototype to test out this technology and demonstrate its possible capabilities. Using last year's system design as a foundation, the team developed a successor with improved transmission specifications. The first part of the process is preparing a file or string of bytes for transmission [1,2,3]. In order to synchronize the transmitter and receiver, the system divides the data into units called "frames", each starting with a preamble to let the receiver know that a transmission has started. The transmitter takes a file, breaks into frames, and inserts preamble sequences before each frame. Then it sends the modified file to a microcontroller unit (MCU) over the serial port. The MCU controls the gate of a transistor based on the data it receives, switching an array of LEDs on when it sees a 1 and turning it off when it sees a 0. This light is picked up by an array of photodiodes on the receiver side. This signal is amplified and filtered to produce a clean signal as similar as possible to what was output by the transmitter MCU. This signal is then sampled by the MCU on the receiver end [4-7]. Each bit sent by the transmitter is sampled 16 times, and the receiver determines whether it's a 0 or 1 based on whichever bit appears more in that 16-bit section (i.e. 14 1s and 2 0s are interpreted as a 1). This data is sent to the computer through a serial connection to be processed by a MATLAB script. The script does the down-sampling and converts the bits into meaningful symbols, either text or audio. At this point, the transmission is complete. However, in our implementation, this success was only apparent on the transmission end of the system. The audio was transmitted flawlessly but the receiver was not able to convert the captured signal back into proper form due to problems in the ADC's of the microcontroller. Sampling errors accumulated to the point where they could not be filtered out by the down-sampling error correction, resulting in a meaningless output [8,9,10].

MATERIALS AND METHODS

The proposed system explains about the transmission of data using two antenna strategies SISO(single input single output) and MIMO(multiple input multiple output) through OSDM. The block diagram describes process of visible light communication .First, The information source produces the data to the modulator where the signal is modulated and the modulated signal is produced to optical driver which has transmitting antennas and transmits the signal through LED and the unwanted signal is filtered and the signal received by the photo detector .the electrical band pass filter receives the signal and filters the electrical signal and the front end amplifier amplifies the signal and the signal produced to the demodulator . The demodulator demodulates the signal and produces to the output unit shown in [Fig. 1].

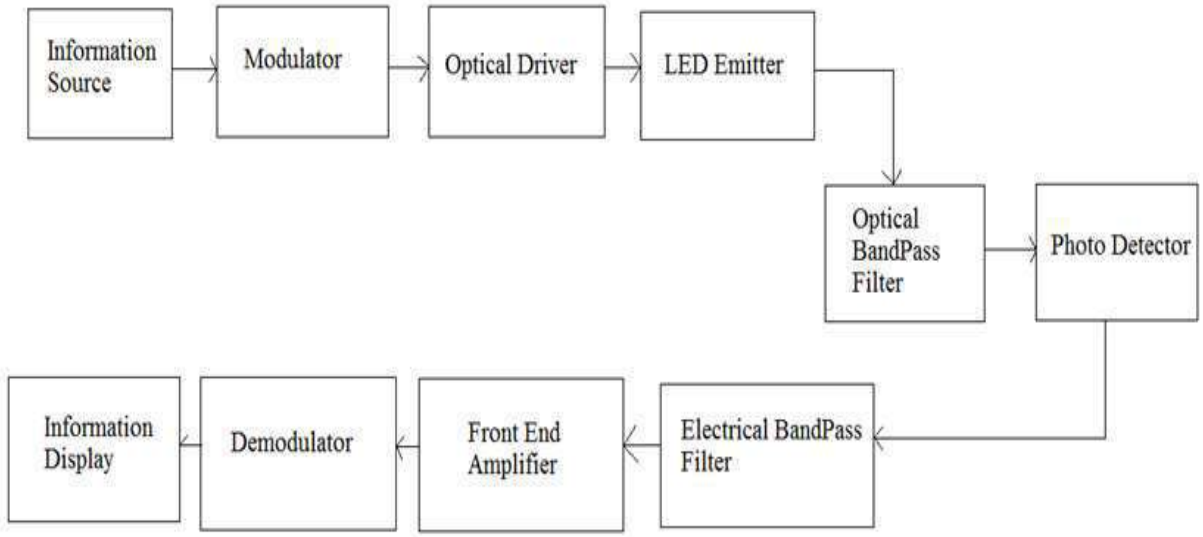


Fig. 1: Demodulator demodulates the signal and produces to the output

Transformer

The purpose of a transmitter is to send data to the receiver so that the other side can process and interpret the data. In our analog design of the transmitter, we used LEDs to transmit light, which would be used to transmit data. As mentioned earlier in the work, the LEDs would modulate on and off to transmit 1s and 0s. The source of the data that would be sent to the LEDs would be the MCU of the transmitter circuit. During the preliminary rounds of building the analog transmitter circuit, it consisted of nothing more than just 20 LEDs and a MCU. While this design worked, it failed to meet one of our goals, which was to transmit at a distance of 1 meter apart to the receiver. The amount of power being sent to the LEDs from the MCU was miniscule. The current from the MCU to the LEDs is not adequate as it was approximately 23 mA of total current coming from the MCU to about 10 LEDs directly with zero resistance. This is approximately comes out to around 2-3 mA for each LED in our system which results in little light produced. In order to alleviate this issue, another power source was needed. We decided that some kind of relay or switch would be needed to handle the switching of the LEDs. We also decided to use the batteries from last year's team to be used as our external power source for the LEDs. In our third and current design as shown in [Fig. 2] below, we decided to change a few things. For the external power source, we decided that four AAA batteries was just not enough for the 10 LEDs so we decided to make use of eight AA batteries. The benefit of using AA batteries is that the power output would last much longer due to having a higher capacity for each battery. Furthermore, more batteries imply a higher voltage, which would lead to an increase in LED brightness as the amount of power for each LED increases. The voltage of our previous battery source is now effectively double in our current build. The resistors between the batteries and the LEDs also were changed as it used to be a higher resistance due to fear of damaging the LEDs with high current but later testing showed that the LEDs appeared to have no issues with a very low resistance. Therefore, the team ended up using two 10Ω resistors in parallel to have the current from the batteries travel through a 5Ω resistance pathway to the 10 LEDs. From our results, the current supplied from the batteries is approximately 360 mA. That would mean about 36 mA is supplied to each LED. The batteries voltage measured to be approximately 11V when the transmitter is off and when under load, the voltages drop down to 10.4V. The final power consumption for each white LED while being supplied with approximately 30 mA is estimated to be around 100 mW so for 10 LEDs, which would be 1.0 W overall shown in [Fig . 2].

Circuit Diagram

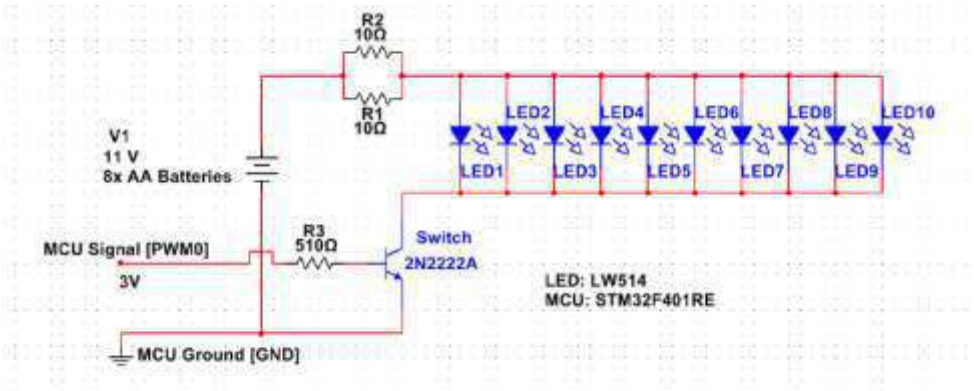


Fig. 2: Transmitter circuit diagram

The other major change to our current design is the transistor used. Originally, a MOSFET was used but was deemed to be unsuitable. This time, we decided to switch to using a BJT. During the testing of transistors, we briefly decided to use the 2N3906 BJT; however, as mentioned in previous sections, it was not adequate. We finally decided to use the 2N2222A BJT, which is very similar to the 2N3906 BJT but is made for slightly higher power usage circuits such as ours. The 2N2222A transistor turns itself on (the resistance between the collector and emitter essentially become 0Ω) when a 0.7 V or higher signal is at coming from the MCU PWM pin. This cutoff voltage is the main reason why we switched over from the IRF520 MOSFET to the 2N2222A BJT as the IRF520 needed 10 V to turn on completely (to lower the resistance between the drain and source to 0Ω). When the BJT turns on, the collector and emitter connect and the power from the batteries reach the LEDs. With the 2N2222A supporting high frequencies, we can turn the switch on and off quickly (up to 150 MHz) so there are no issues with the speed capability.

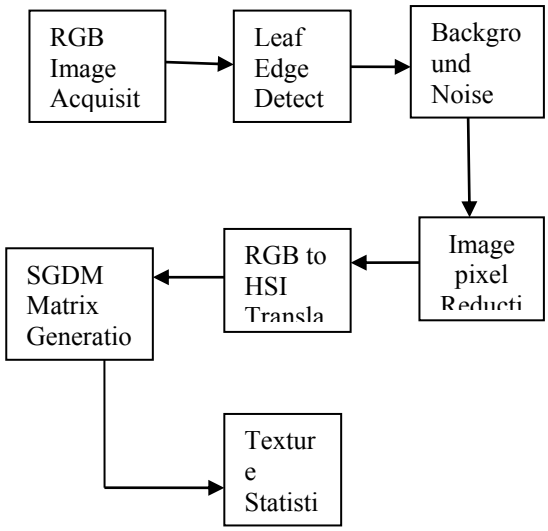


Fig. 2: Image acquisition and image Classification chart

Receiver

The idea behind this circuit was to get a reasonably strong photo current that would only intake the signal from the transmitter LEDs and output that signal to the receiver MCU. To get a strong photo current at up to a foot away, we needed a sizable array of photo diodes. Matching the transmitter, the array had 10 photo diodes to capture the light. With the LEDs shining at 6 inches away, each photo diode produced 1.5 mA of photo current. 10 photo diodes produced 15 mA of photo current. Unlike the transmitter, which uses a resistor to pull a sizeable current from the voltage source through the BJT, the receiver needed a resistor to create a voltage difference with the photo diodes acting as a current source. This resulted in a square wave going from 150 mV to 300 mV, which was smaller than expected. Initially, the idea was to amplify the

signal, and then we could set a voltage cutoff based on the output and there would not be a need for much analog processing beyond that. . We decided upon resistor values of R1 = 1 kΩ and R2 = 10 kΩ to give a gain of 10. This circuit did accomplish the goal of amplifying the input voltage to get a clearer difference between a logic 0 and logic 1, but the problem was that when the output was tied to the ADC input on the MCU, the voltage going in was at a range of about 2.5V to 5V. Not only was the AC component of the signal amplified, but the DC was amplified as well.To deal with this small voltage difference as well as the DC offset, we put this signal through an active high-pass filter. This would accomplish multiple things for our photo diode's output. For one, it would amplify the photo current, making it less likely for the MCU to make a quantization error. Secondly, it would remove the DC offset and allow for more amplification while not running into the problem of hitting the MCU's input ceiling of 3.3V. It is not very useful if the signal goes from 3V-7V if the MCU treats everything above 3.3V as the same.

Removing the offset makes the active range more compatible with the MCU. Third, a high pass filter would filter out all ambient light. Generally, ambient light in any indoor environment would be locked to 60Hz. This noise was actually initially quite apparent in the output and had a significant effect. A high-pass filter got rid of the noise caused by ambient as well as any general white noise. The feedback resistor, in relation to the series resister, determines the amplification of the system. The capacitor, combined with the series resistor, determines the cutoff frequency of the system. With the series resistor set to a value that would create a suitable input voltage (300 kΩ in this case), the capacitor and feedback resistor were chosen to match the desired amplification and cutoff frequency. We wanted to get a gain factor of ~3, and a cutoff frequency between 250 Hz and 500 Hz for the initial tests, where data would be transmitted at 1 kHz but without the 60 Hz ambient light frequencies. Thus, our feedback resistor was set to 1 MΩ and the capacitor was set to 15 nF. Theoretically, the gain should be approximately -3.33, and the frequency cutoff should be about 250 Hz. However, some problems were experienced with the previous design. There were a series of inconsistencies in the output of the previous circuit. At times, it would produce strange output signals. Through a series of tests, it was found that this problem was mostly eliminated by changing how power was supplied to the op-amp. Introducing a duel polarity power supply set up removed the inconsistencies in the output. This is likely because a power supply can more reliably supply power than the batteries could shown in [Fig. 3].

Rectifier Circuit Diagram

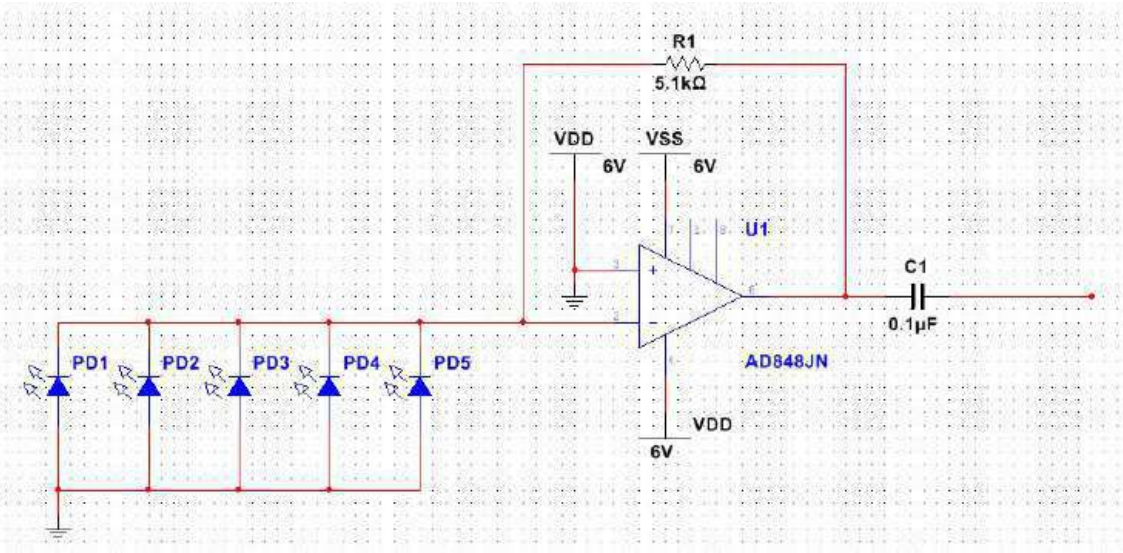


Fig. 3: circuit diagram

RESULTS

Output For MIMO

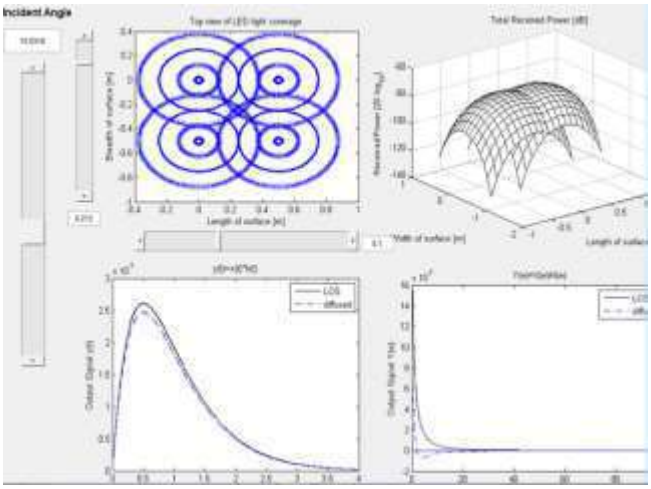


Fig. 4: MIMO Output

Multiple input is multiple output is a antenna strategy which has many inputs and produces many output .The above output describes the top view of LED coverage ,3 dimensional representation of total received power ,width of surface and length of surface. The third plot describes the variation in output signal due to line of sight. As the line of sight has some impact it will reduce the output signal shown in [Fig. 4].

Output For SISO

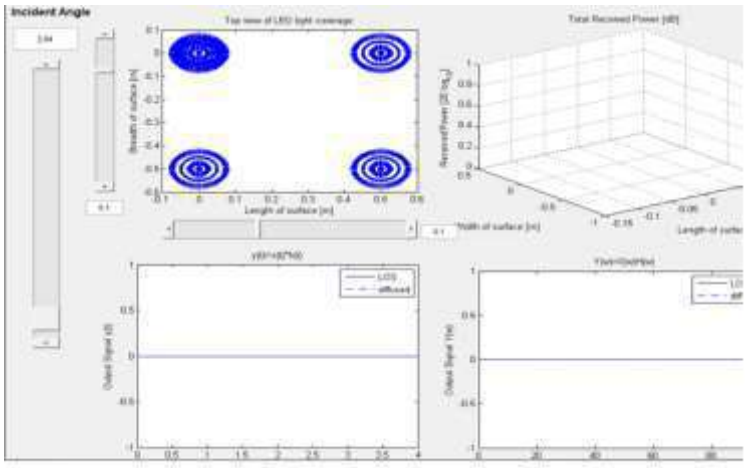


Fig. 5: Output for SISO

Single input is single output is a antenna strategy which has single inputs and produces single output .The above output describes the top view of LED coverage ,3 dimensional representation of total received power ,width of surface and length of surface. The third plot describes the output signal due to line of sight As the line of sight has some impact it will reduce the output signal. As it has single input it doesn't affected by line of sight shown in [Fig. 5].

SNR VS BER(OSDM)

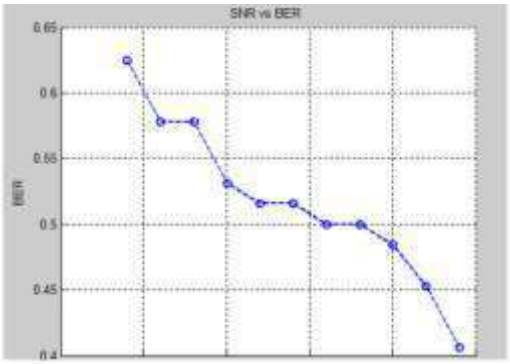


Fig. 6: output graph between SNR vs BER

The above plot describes the SNR VS BER reduced due to the implementation of OSDM shown in [Fig. 6].

CONCLUSION

In VLC communication we used MIMO and SISO based strategies and through this we achieve the communication in indoorwise. In OSDM we have reduced the BER (bit error rate) and increase the SNR (signal to noise ratio). This project was used as a continuation and general improvement of the VLC project completed last year. Utilizing the knowledge the previous team gained through their experiences, a number of improvements could be made. The basic format of these circuit features was overall very similar to the previous year's design. The main difference came in component selection, particularly the photodiodes and the MCU's. The osdm and ofdm combinely used we can make the communication in outdoor also .The signal to noise ratio can be reduced in the future and the bit error rate is also reduced. the equipment cost can be more economical .It can be more commercialized. Loss due to obstacle should be overcome.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Jun Hong Lee, Min Ho Jung, and Jong Ha Shin. [2012] Off-the-Line Primary Side Regulation LED Lamp Driver With Single-Stage PFC and TRIAC Dimming Using LED Forward Voltage and Duty Variation Tracking Control.12.

[2] Ali Mirvakili Department of Electrical and Computer Engineering Tufts University Medford, MA 02155. A Novel Multiple Modes PWM Controller for LEDs. -1-4244-3828-0/09/$25.00 ©2009 IEEE

[3] Krames MR, Shchekin OB, Mueller-Mach R, et al. [2014] Status and future of high-power light-emitting diodes for solid-state lighting [J]. *Journal of Display Technology*. 3(2): 160-175.

[4] Zhang, Member, IEEE, Hulong Zeng, and Ting Jiang. [2012]IEEE 802.15.7 Visible Light Communication: Modulation Schemes and Dimming Support.IEEE Communications Magazine.

[5] Min Ho Jung, and Jong Ha Shin A Primary-Side Control Scheme for High-Power-Factor LED Driver With TRIAC Dimming Capability. IEEE TRANSACTIONS ON POWER ELECTRONICS. 27(11). NOVEMBER 2

[6] AntoBennet M, Suresh R, Mohamed Sulaiman S. [2015] Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", *Journal of Chemical and Pharmaceutical Research*. 07(08): 291-299.

[7] AntoBennet M. [2015] A Novel Effective Refined Histogram For Supervised Texure Classification", *International Journal of Computer & Modern Technology*. 01(02):67-73.

[8] AntoBennet M, Srinath R,Raisha Banu A. [2015] Development of Deblocking Architectures for block artifact reduction in videos, *International Journal of Applied Engineering Research*.10(09): 6985-6991.

[9] AntoBennet M & JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", *Journal of Computer Science*. 8(9):1447-1454.

[10] AntoBennet M & JacobRaglend. [2011] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", *European Journal of Scientific Research*.66 (1):120-129,012.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

DETECTION AND RECONSTRUCTION OF RADAR FOOT PRINTS FOR MAP UPDATING

G.Vijayalakshmi[1], B. Sathyasri [2], V.Mahalakshmi[3] M. Anto Bennet[4*]

Electronics and Communication Engineering, VEL TECH ,Avadi,Chennai 600 062, Tamil Nadu, INDIA

ABSTRACT

The space borne synthetic aperture radar systems acquires imagery with very high spatial resolution, supporting various important application scenarios, such as damage assessment in urban areas after natural disasters. To ensure a reliable, consistent, and fast extraction of the image from the complex synthetic aperture radar scenes. Focusing on the analysis of urban areas, which is of the prime interest of VHR SAR. In this work we proposed a normal method for the automatic detection and 2-D reconstruction of radar foot prints. The method is based on the extraction of a set of low-level features from the images and on their composition to more structured primitives using a production system. The semantic meaning represents the probability that a primitive belongs to a certain scattering class and has been defined in order to compensate for the lack of detectable features in images. The efficiency of the proposed method is demonstrated by processing a 1-m resolution TerraSAR-X spot beam scene containing flat and gable-roof buildings at various settings. The results show that the method has a high overall detection rate and that radar foot prints are well reconstructed, in particular for medium and large buildings.

KEY WORDS
Detection, Radar foot print and MAP Updatation

INTRODUCTION

In the last decade, very high spatial resolution (VHR) space borne remote sensing sensors acquiring data with meter or sub-meter resolutions became widely available. These data have the potential to be employed for various important application scenarios, such as the monitoring of changes in urban areas, the characterization of urban areas, the surveillance of the effects of violent conflicts, and the crisis management after natural disasters. For the latter application scenario, space borne VHR synthetic aperture radar (SAR) sensors, such as Cosmo-SkyMed and TerraSAR-X, are of particular interest, due to their independence on the solar illumination and the relative insensitivity to the weather conditions. One of the main drawbacks of VHR SAR is the complexity of the images, mainly owing to the speckle effect and the side looking geometry of the SAR sensor, hampering the interpretation of the data by non-SAR experts. This is particularly true for urban areas, where the data are mainly characterized by layover, multi bounce, and shadowing effects of the buildings.

Therefore, to support the widespread usage of VHR SAR, robust automatic information extraction methods are essential. Different techniques for building detection and reconstruction from VHR SAR images have been presented in literature. The developed method uses a tree of hypotheses, which is simplified according to a set of semantic rules. The approach is based on the detection of edges and their combination to building footprints. A method for the extraction of buildings and the estimation of their height from stereoscopic airborne radar images while a building extraction method using dual aspect SAR data was presented. An algorithm for building reconstruction from multi-aspect polarimetric SAR (PolSAR) images. The polarimetric information is exploited by employing an edge detector effective on polarimetric images. The retrieved edges are then parameterized by means of the Hough transform to generate the building footprint hypotheses. Another method based on shadow analysis which exploits InSAR data and is suitable for high or isolated buildings. A building detection method using an orthophoto and an InSAR image based on conditional random Techniques for the 3-D reconstruction of buildings using VHR optical data for the 2-D building footprint reconstruction and a single VHR SAR scene for the building height extraction. All the aforementioned works addressed the problem of building detection and reconstruction in VHR SAR images by relying on the availability of ancillary or multi-sensor data, PolSAR, InSAR, or multidimensional airborne data which implies that the area under investigation is imaged more than once with different viewing configurations. This represents a limitation for application scenarios with stringent timing restrictions that do not allow the acquisition of multidimensional SAR data. For these reasons, research on the detection and extraction of buildings from single VHR SAR data is important. This method fails in the detection of buildings if they do not show L-shaped returns. Moreover, it considers only bright lines and discards other relevant features, such as bright areas and shadows. We propose a novel method for the detection and reconstruction of building radar footprints from detected VHR SAR images. Unlike most of the literature methods, it can be applied to single images. Moreover, it is suitable to be used with data acquired by currently operational space borne SAR sensors. In this context, radar footprint refers to the characteristic scattering signature of buildings in SAR. The method integrates the concepts of basic feature extraction and their composition to more structured primitives using a production system. In order to compensate for the lack of detectable features in single images, the concept of semantic meaning of the primitives is introduced and used to generate building candidates and reconstruct radar footprints. The semantic meaning represents the probability that a primitive belongs to a certain scattering class and allows the selection of the most reliable primitives and footprint hypothesis on the basis of fuzzy

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

membership grades. In addition, as shown later in the project, the method is suited to the implementation on computer clusters, thereby making it possible almost-real-time applications. Since their approach relies on sub meter-resolution InSAR data, the hypotheses are based on different information compared to ours. Moreover, we introduce a way to quantitatively evaluate the hypotheses to automatically select the best one, however, this method is based on a global MAP estimation using Monte Carlo methods, while the approach proposed in this project exploits also the shadow information and introduces the concept of semantic meaning and membership grade for each primitive and footprint hypothesis. Moreover, such a work was intended as a tool for the investigation of the limits and merits of information extraction from single images, and was not optimized for building reconstruction purposes. The radar footprint map extracted with the proposed method can be used to derive different information, such as the buildup presence index.

LITERATURE SURVEY

This paper proposes a novel method for the detection and reconstruction of building radar footprints from detected VHR SAR images. Unlike most of the literature methods, it can be applied to single images. Moreover, it is suitable to be used with data acquired by currently operational space borne SAR sensors. In this context, radar footprint refers to the characteristic scattering signature of buildings in SAR. The method integrates the concepts of basic feature extraction and their composition to more structured primitives using a production system. In order to compensate for the lack of detectable features in single images, the concept of semantic meaning of the primitives is introduced and used to generate building candidates and reconstruct radar footprints.The efficiency of the proposed method is demonstrated by processing a 1-m resolution TerraSAR-X spotbeam scene containing flatand gable-roof buildings at various settings.The results show that the method has a high overall detection rate and that radar footprints are well reconstructed, in particular for medium and large buildings[1,5].

Imaging radar has recently become available at geometric resolutions high enough to begin thinking about mapping of individual buildings with their 3-dimensional shapes. This is based on interferometric elevation data as well as magnitude images with their shadows and layover representations from vertical objects. We discuss how an automated procedure can build models of buildings at an accuracy of 1 meter, given radar images with pixel diameters of 30 cms [2,6].

2.5D building modelling is of great interest for visualization, simulation and monitoring purposes in different cases like city and military mission planning or rapid disaster assessment. Manual and automatic extraction of building information from SAR interferometry is complicated by layover, foreshortening, shadowing and multi-bounce scattering, especially in dense urban areas with tall buildings. This can be overcome by using optical imagery to provide the shapes of the building footprints whereas SAR data is used for the refinement of the detected footprints and for providing building height information. In conclusion, building height from SAR interferometry can improve building footprint extraction from optical images and can be combined with the extracted footprints to generate a 2.5D building model [3,7].

In this project we present a novel method for automatic building detection, which also reconstructs the 2D radar footprint of the detected buildings. The method is based on the extraction of a set of low-level features from the images and on their combination in more structured primitives. Then the semantic meaning of primitives is used for the definition of building candidates and for the radar footprint reconstruction. In order to process large VHR SAR images, the method has been implemented on a computer cluster. We demonstrate the effectiveness of the method using a large TerraSAR-X spotlight scene [4, 8].

Proposed method

The proposed technique for the automatic detection and reconstruction of building radar footprints from single VHR SAR images is suited for meter-resolution data.The radar footprints corresponding to very tall buildings have a high probability to be detected.The algorithm does not require the buildings to be isolated, such buildings usually show a clear shadow feature, which is exploited by the algorithm to improve the detection performance.

Modeling of building footprints

The key characteristics of buildings in SAR are the layover, double-bounce, and shadowing effects which are caused by the side-looking and ranging properties of SAR sensors. To illustrate this schematic view of the scattering profile of a simplified flat-roof building model shown in [Fig. 1].

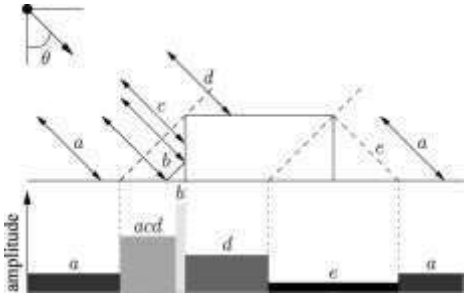


Fig. 1: Scattering model for a flat-roof building with viewing directionfrom left.

The building in the middle, which is modeled as a rectangular box, is imaged by a sensor with incidence angle θ. The annotations a refer to backscattering from the ground surface surrounding the building. acd denotes the layover area where scattering from the ground, from the vertical building front wall and from parts of the flat roof are superimposed since these parts have the same distance to the sensor. The vertical front wall and the surface area in front of the building compose a corner reflector resulting in the bright double-bounce effect b. The scattering area that is only characterized by scattering from the roof is denoted by d. The elevated building occludes parts of the surface behind the building from the radar beam, resulting in the shadow area e. This backscattering profile is flexible with respect to a number of parameters. For instance, for very high buildings, there is typically no area d as the part of the roof is entirely included in the layover area. In fact, both the layover and the shadow areas of the footprint are partially masked by the trees that surround the building. For gable-roof buildings, the theoretic scattering signature is slightly different. The signature has a second bright scattering feature acd at the sensor close side resulting from direct backscattering from the roof. The extent and the strength of this feature depend on the relationship between θ and the roof inclination angle α. For α = θ, the strength of this feature is maximum, whereas its extent is minimum. Moreover, we found that in actual 1-m-resolution TerraSAR-X and Cosmo-SkyMed data, this second bright scattering area is also detectable for buildings with a high aspect angle. Where we show actual scattering signatures from gable-roof buildings with small and large aspect angles, respectively.

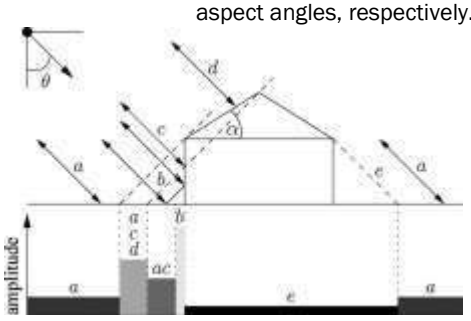


Fig. 2: Scattering model for a gable-roof building with viewing direction from left.

The double-bounce feature is very pronounced. A detailed analysis of the characteristics of the double bounce of buildings with actual TerraSAR-X data and theoretic electromagnetic scattering models presented in showed that this feature has a significant dependency on the building aspect angle. The double bounce has a strong signature for buildings with low aspect angles. Then, it decays significantly in a narrow range of aspect angles, while it drops moderately for larger aspect angles. The method presented in this work will take into account this nonlinear relationship between the strength of the double bounce and the aspect angle shown in [Fig. 2].

RESULTS AND DISCUSSION

Preprocessing and feature extraction

In the preprocessing, the input image is first radiometrically calibrated. Although this step is not strictly necessary, it permits to define the algorithm parameters to be used with SAR images of different data sets and data products acquired by either the same or different sensors. Afterwards, the image is filtered with a Gamma MAP filter in order to reduce the signal variability due to speckle. Both the unfiltered and filtered images are used by the algorithm. The basic features composing building radar footprints in VHR SAR images are extracted from the calibrated image. According to the aforementioned assumptions on building shapes, these are bright linear features with different thicknesses, and dark areas. The former

are usually related to double-bounce scattering or, as the line thickness increases, to layover areas, where the roof or the facade scattering may be dominant depending on the building characteristics. The latter are due to building shadows and low-return areas. These features are sufficient to describe the main parts of a building radar footprint in meter-resolution images shown in [Fig. 3].

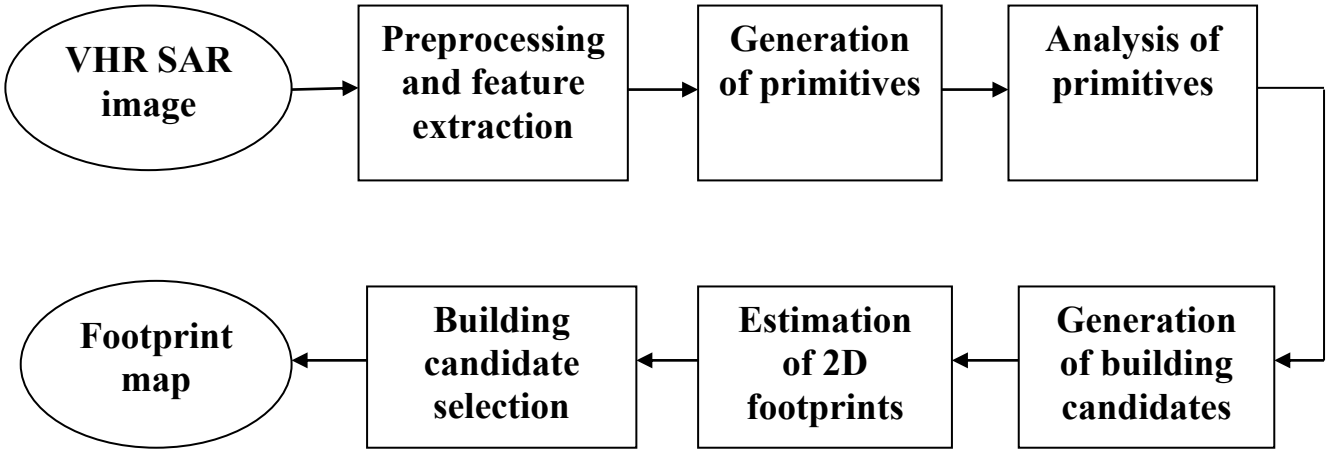


Fig. 3: Block scheme of the processing chain of the reconstruction of building radar footprints in single VHR SAR images.

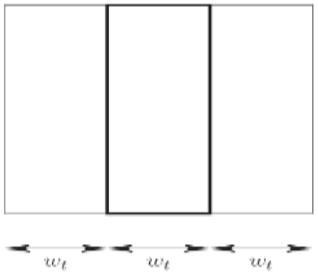


Fig. 4: Definition of the window used by line detector.

Extraction of bright linear features

The extraction of bright linear features is performed on the unfiltered image by means of the line detector. This detector is based on a three-region sliding-window approach and is a well-known algorithm specifically developed for SAR images. In this project, we use as reference for the window size the dimension of the central region, and assume that the lateral regions have the same width and length. The length has been set to ten times the resolution of the image, and 16 directions have been considered for the window. As we are interested in both thin and thick linear features, the detector is applied T times with different increasing window sizes shown in [Fig. 4].

$$Wt\ (t = 1, \ldots, T) \qquad [1]$$

Each filtering is performed independently. The result of each filtering is a detection map, which is then threshold, obtaining binary linear regions which thickness is related to wt. Such regions are vectorized using a rectangular approximation. This is performed by approximating the region skeletons with lines and using such lines as the axis of rectangles of width wt. An example of the detection on a meter-resolution SAR image of an urban area using wt = 5 m. The intermediate results are also shown. For each rectangle r, the local contrast Value Cr is calculated on the filtered image as shown in [Fig. 5].

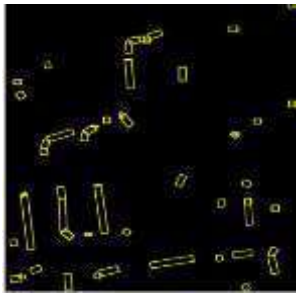


Fig. 5: Meter-resolution TerraSAR-X image of an urban area. **Fig. 6:** Result of the line detection using wt = 5 m.

As a result of the T filtering, we obtain T vector maps containing rectangles corresponding to bright linear features with different thicknesses. These maps are thus merged in one map. It is possible that the same real bright objects are detected independently for different wt, resulting in overlapping rectangles in the merged map shown in [Fig. 6]. In order to reduce the number of rectangles, a down selection step is performed by means of a production net shown in [Fig. 7, 8]. For each combination of two rectangles (i, j), the net tests the following conditions:

1) The width of the two rectangles is similar, and
2) The two rectangles overlap.

Condition 1) is met when

$$|wi - wj| < \delta wmax[2]$$

where wi and wj are the widths of the rectangles, and δwmax is a user-defined threshold.

Condition 2) is fulfilled when

$$A\cap > Ai \cdot At \wedge A\cap > Aj \cdot At$$

where Ai and Aj are the areas of the rectangles, A∩ = Ai ∩ Aj.

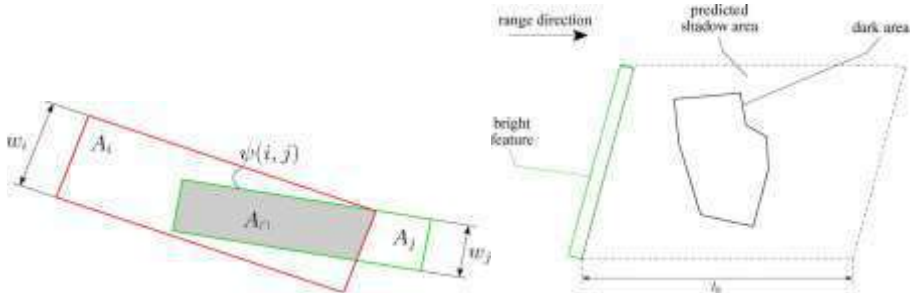


Fig. 7: Measures involved in the rectangle down selection **Fig. 8:** Predicted shadow area and selected dark area

Extraction of dark areas

Dark areas are extracted from the unfiltered image by means of mean shift clustering followed by a threshold operation. This operation selects only the clusters with amplitude values lower than a user-defined threshold. The extracted clusters are then vector zed, and a simplification procedure is applied in order to reduce the number of vertexes describing their shape. Such simplification is not strictly necessary, but it allows the algorithm to work with simpler objects reducing the needed amount of memory. In order to select only the dark regions which are likely to be related to building shadows, the algorithm removes the regions which are not located in the sensor-far side of any bright linear feature. This is done by keeping only the dark areas which overlap with the predicted shadow area of the bright features. The predicted shadow area is determined by taking into account the viewing configuration of the SAR. The maximum range size ls of the expected shadow area is set by the user. The parameters of the mean shift clustering and the value of xs have to be selected by analyzing the amplitude of sample pixels belonging to shadow regions in the SAR image.

Generation of primitives

The goal of this step is to generate the primitives that will be used in the following steps as basis for the composition of building radar footprint hypotheses. Starting from the set of simple extracted bright linear features and dark areas, the algorithm merges adjacent features in order to compose bigger objects. This is done by a production system applied to the vector domain, after a conversion from slant range to ground range, and is aimed at compensating for errors in the feature extraction step shown in [Fig. 9]. The conversion from slant to ground range allows us to define the parameters of the method in the ground domain, which is independent on the incidence angle and thus simpler to handle for an end-user. After their generation, composed objects are given as input to the production system. Therefore, multiple compositions with other simple or composed objects are possible. The set of objects and productions involved in the generation of primitives. The composition of dark areas is described by the production rules P1 and P2. Such rules merge two dark areas when these are adjacent. In this project, the merging is carried out by calculating the convex hull including the two original features.

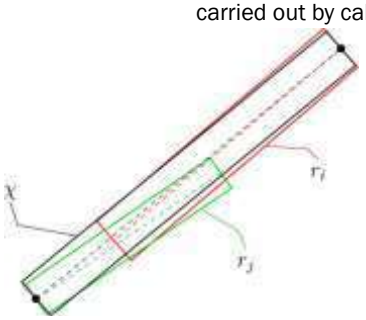


Fig. 9: The merging of two rectangles.

For the case of bright linear features, merged features are generated as new rectangles that have as principal axis the conjunction of the two extremes of the principal axes of the original features which have the largest relative distance.

The algorithm merges two bright features when the following conditions are fulfilled:

1) The features have similar widths,

2) Their orientation is approximately the same,

3) The composed object has an orientation that is approximately the same of the original features.

Condition 1) is equivalent to (2). Condition 2) is fulfilled when

where ψ(i, j) is the angle between the two linear bright features represented by the rectangles i and j and δψmax is user-defined and indicates the maximum angle allowed between two features for which they are considered parallel. The value of δψmax should be on the order of 20◦.

Condition 3) is satisfied when

where χ is the rectangle corresponding to the composed bright linear feature. It is probable that in this step, many bright primitives are generated. In order to reduce their number, a selection procedure as the one described in the previous subsection for bright linear features can be applied. At the end of this step, for the whole set of simple and composed objects, the algorithm stores a set of attributes regarding their size and position, and the amplitude features of the composing pixels.

Analysis of primitives

This step aims at evaluating the semantic meaning of the primitives. Here, we use the term semantic meaning to describe the membership grade of a certain primitive to belong to a predefined scattering class. Different scattering classes are related to different parts of building radar footprints. The choice of the set of semantic classes is related to the types of features extracted from the image and, thus, to the image resolution. For bright primitives, we define four semantic classes: general line, double bounce, roof, and facade. For dark primitives, only the class shadow has been defined. The membership grade of each primitive to belong to a certain semantic class is calculated on the filtered image according to membership functions (MFs) derived empirically for each semantic class.

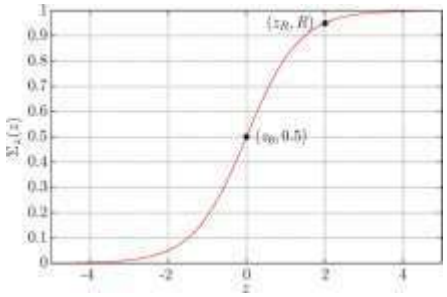


Fig. 10: Sigmoid function Σz(z) defined

The MFs are defined as a product of sigmoid functions. Each sigmoid factor depends on a specific attribute of the primitives. The function Σz(z) gives values in the range (0, 1). For each sigmoid function, two parameters need to be specified: the value of z for which the sigmoid returns a high likelihood R(zR), and the value corresponding to the center of the sigmoid (z0), implicitly setting the slope of the function. The MFs, which relate bright primitives to the relative semantic classes, are defined according to the tree. The number of sigmoid functions composing the MF for a semantic class is smaller or equal to the number of branches that connect the root to the final leaf. In the following, we describe in detail the MFs of each semantic class for both bright and dark primitives, by also suggesting the range of parameters that is most suited for the related scattering class. Unless otherwise stated, such values have been estimated by analyzing the scattering properties of a set of samples of the considered scattering classes manually selected on the meter resolution TerraSAR-X input images used in this paper shown in [Fig. 10]. As the images are calibrated, the suggested values related to pixel amplitude can be considered generally valid. In the case of images acquired at different resolution and or with a sensor with different characteristics, some of the values should be estimated again.

Bright primitives
General line

The membership grade of a primitive to the class general line depends only on its width. It gives a measure of the membership of the primitive to the high-level class thin line, which depends on the primitive width w.

Double bounce

The double-bounce effect appears in VHR SAR images as relatively thin bright lines. It is more evident when the building wall is parallel to the azimuth direction, i.e., its aspect angle is close to zero. In such a case, the MFs of the classes general line and double bounce give very similar values. Such values have been chosen according to our previous studies about the double bounce effect in VHR SAR images.

Roof

The class roof is the most specific, as it appears as leaf for every branch combination. This is due to the intrinsic uncertainty given by the fact that we are using only one VHR SAR image and that we are considering meter-resolution images. Indeed, the signature of a building roof could be either a thin line or a homogeneous rectangular area, or a non-homogeneous rectangular area. Therefore, for the class roof, the final membership grade is calculated as the maximum of the membership grades given by the three MFs corresponding to the three occurrences of the class in the tree. These refer to the homogeneity of the pixels contained in the primitive. The homogeneity is measured using as parameter the coefficient of variation σ of the pixels.

Facade

As reported in the tree, the semantic class facade includes primitives with a relevant width and which pixels have non-homogeneous values.This is the general scattering behavior of building facades, where returns coming from structures like windows or balconies give a strong textured signature in the radar footprint. As a further constraint, the aspect angle of the building should not be too high. Indeed, the facade scattering area in the radar footprint becomes smaller with increasing aspect angles shown in [Fig. 11].

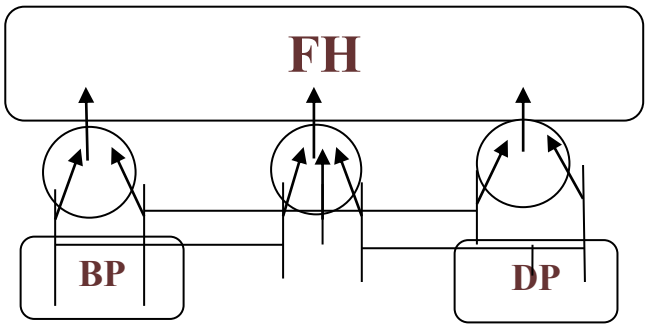


Fig.11: Production net for the generation of building radar footprint hypotheses

Dark primitives

For dark primitives, only the semantic class shadow has been defined. The MF of this class takes into account the mean and the coefficient of variation of the pixels contained in the primitive. The MF is tuned in order to penalize dark primitives with high mean value and high coefficient of variation.

Two-dimensional reconstruction

The 2-D radar footprint reconstruction aims at refining the detection of both, the bright part and the dark part (if present) of the footprint hypotheses selected in the previous step. This is performed in order to reduce the effect of imprecision coming from the feature extraction and primitive generation steps, and to provide reliable outputs which can be used as a starting point to estimate parameters of the buildings, such as their length, width, and height (with the limitations imposed by the fact that only a single image is available). The result of this procedure is thus the final map of the building radar footprints detected and reconstructed from the input VHR SAR image. As a first step, the algorithm generates for each footprint hypothesis a best-fit rectangle which includes its bright primitives. If only one bright primitive is present, the best-fit rectangle and the bright primitive match. Then, the rectangle is translated, rotated, expanded, and shrunk with the goal to maximize Cr. The maximization is carried out using a particle swarm optimization approach, which is a well-known iterative method suited for the optimization of problems without a priori assumptions. A similar approach was applied in for binary images and using a different optimization strategy. The rectangles which become smaller than the minimum sizes set in the previous steps of the algorithm are deleted. Moreover, it is possible that some rectangles move and overlap. Therefore, the algorithm deletes overlapping rectangles, and thus the corresponding footprint hypotheses, keeping only the rectangles associated to the hypotheses with the highest scores Sh. A refinement procedure is carried out also for the dark part of the footprint hypothesis, when it is present. In fact, a good knowledge of the size of the shadow area of a building can be exploited for the retrieval of the building height. The refinement aims at expanding the dark primitive on pixels with amplitude values similar to those of shadows in the sensor far side area of the reconstructed bright primitives. To this end, the center of the dark primitive is used as seed for a region growing algorithm which, starting from an initial circular contour, stretches its border to fit the dark area around the seed. The chosen implementation is a level-set algorithm which moves the contour by including the pixels which have amplitude values in the range [0, mSR]. The resulting regions are cut in the azimuth direction in order to match the extension of the reconstructed bright part of the footprint hypothesis. Indeed, the reconstructed regions are associated to building shadow areas, which cannot be larger than the corresponding buildings in the azimuth direction. The size of the reconstructed dark areas in the range direction depends only on the radiometric measurements in the image. As the proposed technique uses as input only one VHR SAR image and no a priori information is available, it is not possible to detect the end of the shadow region by other means. This may lead to shadow areas which are longer than real shadows because of low scattering areas behind the buildings (e.g., roads, parking lots). This problem can be partially mitigated by imposing a maximum shadow range size IS set by the user. Shadows longer than IS are cut to IS, and a flag is set to notice the user about the lower reliability of the reconstructed shadow.

EXPERIMENTAL RESULTS

In this section, we show the results obtained by applying the proposed methodology to a real meter-resolution large SAR image. After a brief description of the used data set, we show and analyze qualitatively the results obtained on the whole image following the grid-computing approach described. Then, we focus on two subsets of the image in order to assess quantitatively the accuracy of the method.

Data set description

The effectiveness of the proposed method has been tested on a TerraSAR-X image of the city of Dorsten, Germany. The image has been acquired in HH polarization in spotlight mode, resulting in a geometrical resolution of approximately 1.1 m × 1.2 m (azimuth × slant range). The incidence angle varies between 50.3◦ and 51.0◦. The original scene has been cut to a subset of 2800 × 3712 pixels, covering an area of approximately 10 km2. The cut includes both urban and rural areas. Urban areas are characterized by both flat- and gable-roof buildings at various settings. The SAR test image and an optical image corresponding to the same area taken from Google Maps.

RESULTS ON SUBSET 1

The values of such parameters have been chosen according to the guidelines in table 1. The results obtained are shown in Fig. 12. The method shows in overall a high detection rate. False alarms are mostly related to the scattering from objects different from buildings (e.g., trees, garages) that show radar footprints similar to those of buildings.

Table 1: Parameters used in the feature extraction and primitive generation steps in the experiments carried out with the proposed

Parameter	Value
T	7
w1,…..,w7	3,5,….,15
Wmax	3
At	0.5
Xs	-12.2db
Ls	30m
Ψmax	20°

A particular case is represented by bridges, which have been also detected. Such structures can be easily masked, either using a priori information about the presence of rivers, or by extracting the rivers directly from the SAR. The radar footprints of complex buildings which do not correspond to the rectangular model used in this project are mostly detected with some reconstruction errors. In general, the proposed method detected and reconstructed quite precisely the radar footprints of medium- and big-size buildings that fulfill the rectangular model. Radar footprints of small adjacent buildings aligned in regular patterns are also detected, but in some cases are considered as belonging to a single building. Small buildings which do not show clear features are not detected by the method. However, considering the use of a single SAR image, the results can be considered qualitatively very satisfactory. Moreover, it is worth noting that if the proposed method is applied in order to derive indexes of the presence of buildings, reconstruction errors do not represent a critical issue shown in [Fig. 13].

Fig. 12: Original TerraSAR-X image of the considered area, viewing direction from left

Fig. 13: Reconstructed building radar footprints on the SAR image

Editor: Dr. K. Sakthisudhan

Table2. Algorithm Performance for subset 1, subset 2

	Building size	Number of buildings	Detected	False alarms	Split	Merged
Subset 1	Large	21	19	0	2	1
	Medium	26	22	2	4	3
	Small	66	35	9	1	9
Subset 2	Large	12	12	0	4	0
	Medium	27	23	2	4	3
	Small	53	34	9	1	8
Subset 1+2	Large	33	31	0	6	1
	Medium	53	45	4	8	6
	Small	119	69	18	2	17

RESULTS ON SUBSET 2

This area is characterized by a large number of trees locate dalong the streets. Such trees often mask the radar returns also from medium-sized buildings. Moreover, small buildings are usually quite irregular and show many structures on their walls. This subset is thus a challenging benchmark for the proposed technique. Table reports the results obtained for the subset 2, shows the correct and missed detections on theoptical image. As for the subset 1, the detection rate for the classes large and medium is very good. For the class small, performances are less satisfactory. The number of split buildingsis 1 for the class small, 4 for the class medium, and 4 for the class large; while the number of merged buildings is 8 for the class small, 3 for the class medium, and 0 for the class large. The total number of false alarms is 11. As for subset 1, the most of them are related to small false building radar foot prints. In overall, considering the issues mentioned at the beginning of this paragraph and the limited amount of information used by the proposed technique. The results can be considered very good. In order to provide a more general view of the results obtained by the proposed method, Table II also reports the overall results computed by summing the results of the subsets 1and 2. The total statistic confirms the trend highlighted for the single subsets, i.e., the algorithm has a high detection rate for medium and large buildings, with a limited amount of false alarms, whereas its performance decreases in the case of small buildings, which are associated to most of the total number of false alarms. It is worth noting that it is possible to mitigate this problem by imposing a rule for discarding the footprints smaller than a user-defined minimum footprint size. As a consequence, the number of false alarms would be considerably reduced and the detection of radar footprints of small buildings would not bea target of the method anymore. This is a reasonable strategy to adopt for tuning the proposed technique only on the detection of medium and large buildings.

Selection of algorithm parameters

The tuning of the parameters has been performed according to the scene investigated. However, some parameters are not strictly related to the image analyzed and can be set a priori following general rules. Moreover, many of the considered parameters have a clear physical meaning that helps the user to include its prior knowledge on the scene in the detection algorithm. In addition to the guidelines already provided, in this section, we analyze more in detail the role of the parameters of the proposed method.

Feature extraction and primitive generation

In these steps, the main parameters of the proposed technique are related to the detection and generation of bright rectangles, and to the extraction of the shadows. The possible range of values for the window of the line detector wt should be set between the expected thickness of thin linear features and the maximum size of the buildings which has to be extracted. The sampling of the range of wt, given by the number of filtering T, should assure that most of the linear features can be effectively modeled with the considered values of wt. The minimum value for δwmax has to be greater than the width sampling resulting from the definition of the values of wt. On the one hand, a value smaller than this quantity would not allow the algorithm to down select effectively the rectangles produced in the feature extraction step. Moreover, the procedure for the generation of primitives would combine only rectangles with approximately the same width. On the other hand, a value much greater than the width sampling would make the algorithm to down select too many rectangles, and combine features with much different widths. According to our tests, a good choice for the value of δwmax is 1.5 times the width sampling used in the line detection. Regarding the parameters At and δψmax, high values for At and low angles for δψmax make conditions (3) and (4) too stringent, respectively.

Analysis of primitives

In this step, the main parameters to be set are those related to the MFs defined for the different scattering classes. The choice of the value of R is not critical, and R = 0.999 can be considered as a fixed

value. The parameters wthin R = wthick R , wthin 0 , and wthick 0 used in this project can also be considered general. Indeed, they are given in meters, so that they do not depend on the resolution of the system. According to our tests, setting wthin R = wthick R to a value 2–3 m greater than the expected thickness of the linear signatures due to the double-bounce effect gives the best results, as the procedure which creates rectangles from the output of the line detector may overestimate their actual thickness.

Selection of hypotheses and 2-D radar footprint reconstruction

The parameter Nh is related to the reliability assigned by the user to the footprint hypotheses composed by only two primitives. In our tests, by setting this parameter to higher values resulted in detection maps with less hypotheses composed by three primitives, as expected. Indeed, increasing the value of Nh makes three-primitive hypotheses to have higher probability to score lower than those composed by two primitives. Therefore, three primitive hypotheses have higher probability to be discarded when they overlap with others made up of two primitives. This does not affect significantly the detection rate of the proposed method, but it increases the probability that the extracted footprints are not well-reconstructed (e.g., shadows are missing even though they were detected). On the contrary, by setting Nh to low values would increase the number of missed detections.Therefore, the choice of Nh should be done by the user as atradeoff between reliability of the reconstruction and detectionperformance.The pair of parameters (δdR, δd0) and (δψR, δψ0) are relatedto the vicinity and relative orientation of the primitives, respectively.The values proposed in this project can be considered general for the defined scattering classes. Note that, using these values, the sigmoid functions present are quite smooth, thus mitigating the effect of possible errors in feature extraction. The last parameter to be discussed is Sh, min. This parameter gives the tradeoff between false and missed detections. According to our tests, the use of high Sh, min results in a greater number of missed detections, as expected. However, the number of false alarms is not reduced significantly. Indeed, these are usually related to footprints of other man-made structures, or trees, which actually appear as related to buildings. For this reason, values in the order of 0.6–0.7 are suggested.

Computational load

The test image described in Section V-A has been processed using a cluster composed by 16 AMD Opteron 6172 CPUs, for a total of 192 cores, with 4 GB of RAM per core. The image has been split on tiles of 300 × 300 pixels with an overlapping offset of 30 pixels with the neighbors. The total number of tiles was thus 154, and each tile was processed by one core. The total processing time was about 45 min. With the same infrastructure, it is thus possible to process a whole spotlight image of about 6000 × 10.000 pixels in less than 3 h. We also tested the proposed technique using a smaller cluster composed by eight commercial workstations equipped with Intel Core i7-870 quad-core processors and 8 GB of RAM. The total processing time for the test image on this smaller architecture was about 1 h and 30 min, which is a good performance in terms of operational application of the algorithm.

Input Image

We present a novel method for automatic building detection, which also reconstructs the 2D radar footprint of the detected buildings. The method is based on the extraction of a set of low-level features from the images and on their combination in more structured primitives. In order to process large VHR SAR images, the method has been implemented on a computer cluster. We demonstrate the effectiveness of the method using a large Terra SAR-X spotlight scene. In order to exploit this information in different application scenarios, robust building detection and reconstruction methods are essential. Different techniques for building detection and reconstruction from VHR SAR. They mainly rely on the availability of multi-modal data that require multiple acquisitions with different viewing configurations shown in [Fig. 14].

Fig. 14: Selecting input image.

In the preprocessing, the input image is first radio metrically calibrated. Although this step is not strictly necessary, it permits to define the algorithm parameters to be used with SAR images of different data sets and data products acquired by either the same or different sensors. Afterwards, the image is filtered with a Gamma MAP filter in order to reduce the signal variability due to speckle. Both the unfiltered and filtered images are used by the algorithm. The basic features composing building radar footprints in VHR SAR images are extracted from the calibrated image. According to the aforementioned assumptions on building shapes, these are bright linear features with different thicknesses, and dark areas shown in [Fig. 15].

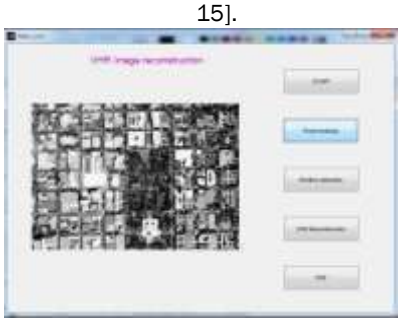


Fig. 15: Preprocessed output image.

This step aims at evaluating the semantic meaning of the primitives. Here, we use the term semantic meaning to describe the membership grade of a certain primitive to belong to a predefined scattering class. Different scattering classes are related to different parts of building radar footprints. The choice of the set of semantic classes is related to the types of features extracted from the image and, thus, to the image resolution. For bright primitives, we define four semantic classes: general line, double bounce, roof, and facade. For dark primitives, only the class shadow has been defined. For dark primitives, only the semantic class shadow has been defined. The MF of this class takes into account the mean and the coefficient of variation of the pixels contained in the primitive. The order in which the bright primitives are aggregated is also taken into account, at least two hypotheses will be generated for each pair of bright primitives shown in [Fig. 16].

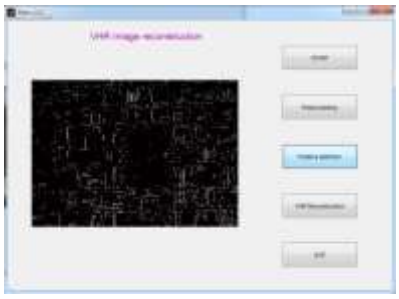


Fig. 16: Primitive detection process.

The 2-D radar footprint reconstruction aims at refining the detection of both, the bright part and the dark part (if present) of the footprint hypotheses selected in the previous step. This is performed in order to reduce the effect of imprecision coming from the feature extraction and primitive generation steps, and to provide reliable outputs which can be used as a starting point to estimate parameters of the buildings, such as their length, width, and height. The result of this procedure is thus the final map of the building radar footprints detected and reconstructed from the input VHR SAR image. The size of the reconstructed dark areas in the range direction depends only on the radiometric measurements in the image. As the proposed technique uses as input only one VHR SAR image and no a priori information is available, it is not possible to detect the end of the shadow region by other means. This may lead to shadow areas which are longer than real shadows because of low scattering areas behind the buildings shown in [Fig. 17].

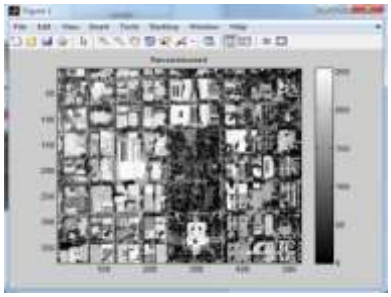


Fig. 17: Reconstructed image.

CONCLUSION

In this paper, the problem of the detection and reconstruction of building radar footprints in VHR SAR images has been addressed. It extends state-of-the-art feature extraction and composition steps to more structured primitives using a production system and by introducing the concept of semantic meaning. This has been done in order to compensate for the lack of information due to the fact that only one VHR SAR image is used as input. It allows the technique to select the most reliable primitives and footprint hypotheses during its processing steps. As a further refinement; the proposed technique also reconstructs the detected radar footprints. The goal of this step is to provide as output a map which can be used as a starting point for further calculations. Moreover, by exploiting the reconstruction of the shadow areas, height retrieval techniques can be also applied to estimate building heights. To make it possible to use the proposed technique on large VHR SAR images in near real time, we also proposed and implemented an infrastructure based on a computer cluster for the processing of large VHR SAR scenes. The system can be programmed in integrated chip and can be inhibited into the camera itself, so that real time detection can be achieved. The threshold value which has been calculated manually in the process based on the approximate median of the object intensity can be automated in future for higher efficiency. More effective threshold techniques can be implemented for low quality videos which would help satellite surveillance applications.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] AUTOMATIC DETECTION AND RECONSTRUCTION OF BUILDING RADAR FOOT PRINTS FROM SINGLE VHR SAR IMAGES " Adamo Ferro, Dominik Brunner, and Lorenzo Bruzzone, IEEE Transactions on geosciences and remote sensing, 61(2), February 2013.

[2] BUILDING RECONSTRUCTION FROM SAR IMAGES AND INTERFEROMETRY" Franz W. Leberl, Regine Bolter, IEEE Transactions on geosciences and remote sensing, May 2012.

[3] RECONSTRUCTION OF BUILDING FROM VHR SAR AND OPTICAL DATA BY USING OBJECT ORIENTED IMAGE ANALYSIS" Namhyum Kim , February 2011.

[4] BUILDING DETECTION AND RADAR FOOTPRINTS RECONSTRUCTION FROM SINGLE VHR SAR IMAGES" Adamo Ferro, Dominik Brunner, and Lorenzo Bruzzone, March 2011.

[5] AntoBennet M, JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images, Journal of Computer Science, 8(9):1447-1454,

[6] AntoBennet M, JacobRaglend. [2011]Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", European Journal of Scientific Research, 66 (1):120-129

[7] AntoBennet M , Resmi R. Nair, Mahalakshmi V, Janakiraman G. [2016] Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography, Indian Journal of Science and Technology, 9(2):1-7

[8] AntoBennet M, JacobRaglend. [2012] 'A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach', Journal of Applied Sciences Research, 8(5):2429-2438.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

AUTOMATED LOCALIZATION OF OPTIC DISC IN RETINAL VASCULAR CHANGES AND DIABETIC RETINOPATHY

M Anto Bennet[1*], B Thamilvalluvan[2], S Yacoop Rahman[3], M Venkatesh[4], J Albert Santhosh Raj[5]
Dept of Electronics and Communication Engineering, VEL TECH, Avadi, Chennai 600 062, TN, INDIA

ABSTRACT

Glaucoma is one of the most common causes of blindness. The number of people having severe vision loss in developing countries. Robust mass screening may help to extend the symptom-free life for affected patients. To realize mass screening requires a cost-effective glaucoma detection method which integrates well with digital medical image processing. To address these requirements, the proposed novel low cost automated glaucoma diagnosis system based on hybrid feature extraction from digital fundus images. This paper discusses a system for the automated identification of normal and glaucoma classes using Higher Order Spectra (HOS), Trace Transform (TT), and Discrete Wavelet Transform (DWT) features. The same features with SVM classifier produce a less accuracy. So the proposed extracted features are fed to a Expectation Maximization (EM) classifier to achieve a good result. This was able to identify glaucoma and normal images. Furthermore, we propose a novel integrated index called Glaucoma Risk Index (GRI) which is composed from HOS, TT, and DWT features, to diagnose the unknown class using a single feature. Hence this GRI will aid clinicians to make a faster glaucoma diagnosis during the mass screening of normal/glaucoma images.

KEY WORDS
Discrete Wavelet Transform (DWT), Expectation Maximization(EM), Glaucoma Risk Index

Received: 24 October 2016
Accepted: 20 December 2016
Published: 15 February 2017

***Corresponding Author**
Email:
bennetmab@gmail.com

INTRODUCTION

The human eye is a complex biological device. The mechanism of cameras often compared with the working of the eye, as shown in [Fig.1]. Light entering the eye is first refracted when it passes through the cornea. It then passes through the pupil and is further refracted by the lens. Finally, it reaches the retina and is converted to electrical signals by photosensitive photoreceptor. The electrical signals are transmitted to the brain along the optic nerve. The cornea is the transparent front part of the eye. It is the first structure that is able to refract the light entering the eye. However, the focal distance of the cornea is fixed, which means that the cornea can only refract light with a constant angle. The lens, on the other hand, can adjust its focal distance so that incoming light can be focused on the retina. The lens is a transparent structure lying behind the iris and the pupil. The iris is a membrane organ in the eye. It controls the diameter and size of the pupil and hence the amount of light reaching the retina. The movement of iris is controlled by the iris dilator muscle. The pupil is an opening in the center of the iris. It allows light to enter the eye and reaches the lens. The pupil appears to be black because most of the light entering the pupil is absorbed. The vitreous is the transparent, colorless, gelatinous mass that fills the space between the lens and the retina. It is also referred to as the vitreous body or vitreous humor. Unlike the fluid in the frontal part of the eye which is continuously replenished, the gel in the vitreous is stagnant. So if blood or cytosol gets into the vitreous, they may not be reabsorbed for an extended period of time. A vitreous hemorrhage is a typical symptom of diabetic retinopathy. The optic disc or optic nerve head is the location where ganglion cell axons exit the eye. The optic nerve is a bundle of more than one million nerve fibers. The optic nerve connects the retina to the brain. It is also the place where all retinal blood vessels originate and converge. The optic disc is placed 3-4mm to the nasal side of the fovea [1,2,3].

Retinal Images

The retina is a layered tissue lining the interior of the eye that enables the conversion of incoming light into a neural signal that is suitable for further processing in the visual cortex of the brain. It is thus an extension of the brain. The ability to image the retina and develop techniques for analyzing the images is of great interest. As its function requires the retina to see the outside world, the involved ocular structures have to be optically transparent for image formation. Thus, with proper techniques, the retina is visible from the outside, making the retinal tissue, and thereby brain tissue, accessible for imaging noninvasively. Because the retina's function makes it a highly metabolically active tissue with a double blood supply, the retina allows direct noninvasive observation of the circulation shown in [Fig.2].Thus, because of its architecture dictated by its function both diseases of the eye, as well as diseases that affect the circulation and the brain can manifest themselves in the retina. These include ocular diseases, such as macular degeneration and glaucoma, the first and third most important causes of blindness in the developed world. A number of systemic diseases also affect the retina. Complications of such systemic diseases include diabetic retinopathy from diabetes, the second most common cause of blindness in the developed world, hypertensive retinopathy from cardiovascular disease, and multiple sclerosis. Thus, on the one hand, the retina is vulnerable to organ-specific and systemic diseases, while on the other hand, imaging the retina allows diseases of the eye proper, as well as complications of diabetes, hypertension and other cardiovascular diseases, to be detected, diagnosed and managed. This review focuses on quantitative approaches to retinal image analysis. Principles of 2-D and 3-D retinal imaging are outlined first. Special

emphasis is given to fundus and optical coherence tomography (OCT) image analysis and its use to provide comprehensive descriptions of retinal morphology and function. The described methods cover the developments of the past decade and were selected with respect to their potential for screening-motivated computer-aided detection of retinal abnormalities as well as for translational clinical applications including improved retinal disease diagnoses and image-guided retinal therapy. As such, the methods presented are expected to influence routine clinical patient care in the years to come [4,5].

Disease Specific Analysis of Retinal Images

The everyday cost associated with eye care providers' decisions and the ever-increasing numbers of retinal images to be reviewed are the major motivations for the adoption of image analysis in ophthalmology. Clearly, since clinicians are costly experts, they need to optimize the time devoted to each patient, whether their cost is born by patients, third party insurers, or society as a whole. As presented in the following sections, the development of new imaging technology invariably results in rapidly increasing amounts of data collected as part of any specific retinal imaging exam. The amount of information provided by the current generation of scanners and cameras is already exceeding the limit of clinicians' ability to fully utilize it. When factoring in that clinicians are subjective, and their decisions suffer from the inter- and intra-observer variability, the need for reliable computerized approaches to retinal image analysis is more than obvious, if for no other reason, than to increase the precision with which patients are managed. An additional important reason for incorporating automated analyses of retinal images in patient management is the potential societal benefit of increasing clinician productivity in a routine population screening setting. While the patient management decision making and population screening scenarios are somewhat different and specific, they both require quantitative retinal image analysis to be rapidly translated to everyday use [6,7].

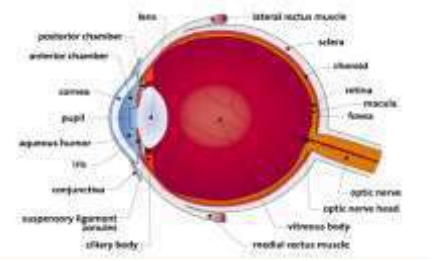


Fig. 1: Human Eye Structure

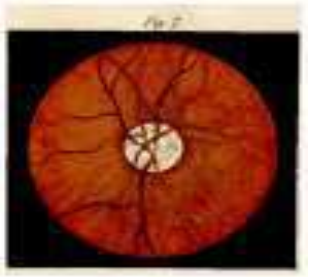


Fig. 2: Image of human retina

Retinal Camera & Photograph

A camera attached to an indirect ophthalmoscope aimed at photographing the image of the fundus of the eye. This image is produced by the objective of the ophthalmoscope at the first focal point of the objective of the viewing microscope (and of the camera), which forms an image on the film. A flip mirror within the optical path of the viewing microscope allows the observer to view the image of the fundus and focus it, thus ensuring that the image being photographed is as clear as that being viewed. Fundus cameras usually require a dilated pupil of about 4 mm and their fields of view extend up to 45°. They provide an objective photographic record of any condition in the fundus. They can also be used to take photographs of the anterior segment of the eye.

Currently, regular screenings are conducted and retinal images are obtained using fundus camera. However, a large amount of images are obtained from these screenings and it requires trained ophthalmologists to spend a lot of time for manual analysis and diagnosis. Hence, automatic detection is desired as it can help to improve productivity and be more cost effective shown in [Fig.3 &4].

Fig. 3: Retinal camera

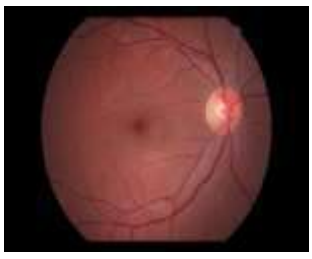


Fig. 4: Normal eye retinal photograph

Glaucoma

Glaucoma is a condition that causes damage to eye's optic nerve and gets worse over time. It's often associated with a buildup of pressure inside the eye. Glaucoma tends to be inherited and may not show up until later in life. The increased pressure, called intraocular pressure, can damage the optic nerve, which transmits images to the brain. If damage to the optic nerve from high eye pressure continues, glaucoma will cause permanent loss of vision. Without treatment, glaucoma can cause total permanent blindness within a few years. Glaucoma usually occurs when pressure increases in eye shown in [Fig.5&6]. This can happen when eye fluid isn't circulating normally in the front part of the eye. Normally, this fluid, called aqueous humor, flows out of the eye through a mesh-like channel. If this channel becomes blocked, fluid builds up, causing glaucoma. The direct cause of this blockage is unknown, but doctors do know that it can be inherited, meaning it is passed from parents to children .Less common causes of glaucoma include a blunt or chemical injury to the eye, severe eye infection, blockage of blood vessels in the eye, inflammatory conditions of the eye, and occasionally eye surgery to correct another condition. Glaucoma occurs in both eyes, but it may involve each eye to a different extent. Open-angle glaucoma. Also called wide-angle glaucoma, this is the most common type of glaucoma. The structures of the eye appear normal, but fluid in the eye does not flow properly through the drain of the eye, called the trabecular meshwork. Angle-closure glaucoma. Also called acute or chronic angle-closure or narrow-angle glaucoma, this type of glaucoma is less common but can cause a sudden buildup of pressure in the eye. Drainage may be poor because the angle between the iris and the cornea is too narrow.To diagnose glaucoma, an eye doctor will test your vision and examine your eyes through dilated pupils. The eye exam typically focuses on the optic nerve which has a particular appearance in glaucoma. In fact, photographs of the optic nerve can also be helpful to follow over time as the optic nerve appearance changes as glaucoma progresses. Glaucoma tests are painless and take very little time.

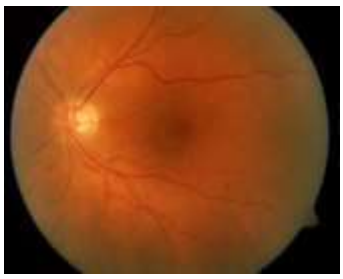


Fig. 5: Normal Image

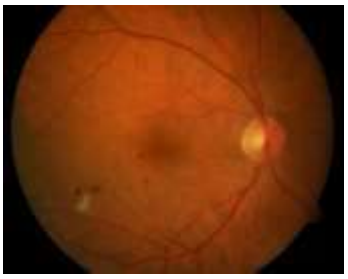


Fig. 6: Abnormal Image

MATERIALS AND METHODS

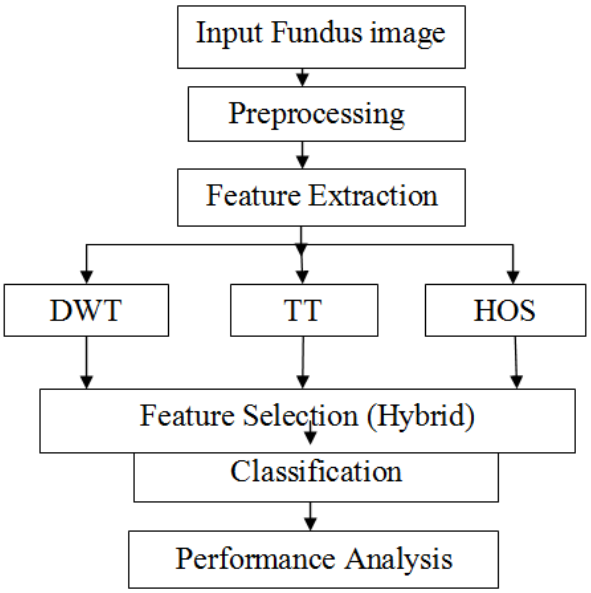


Fig. 7: Block Diagram of Glaucoma Based on Hybrid Features in Retinal Images

Input Image

The input image is the retinal fundus eye image which is RGB image get from the ophthalmologists. The input image taken in the form of JPEG, PNG, or BITMAP format. In a retinal image, the optic disc occupies a small area of the entire retinal image. Retinal images need to be preprocessed before the feature extraction is shown in [Fig.8].

Fig. 8: Input Image

Preprocessing

Preprocessing of retinal images is first step in the automatic diagnosis of retinal disease. The quality of retinal image is not good, so it is necessary to improve the quality of retinal image the purpose of preprocessing is to improve the noisy area from retinal image shown in [Fig.7].

Histogram Equalization

Enhancing the fundus image contrast will aid the feature extraction process. In this work, colored (RGB) eye images are converted to gray scale image by forming weighted sum of R, G, and B.

I gray=0.2989*R+0.5870*G+0.1140*B

After the conversion of gray level, the histogram equalization is done to improve the quality of input images shown in [Fig.9&10].

Preprocessing for Normal Image

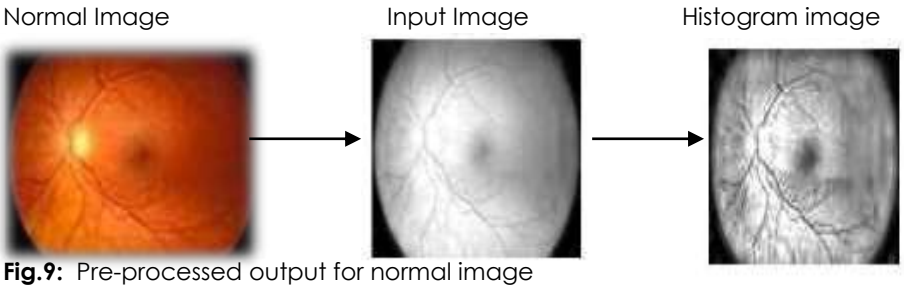


Fig.9: Pre-processed output for normal image

Preprocessing for abnormal Image

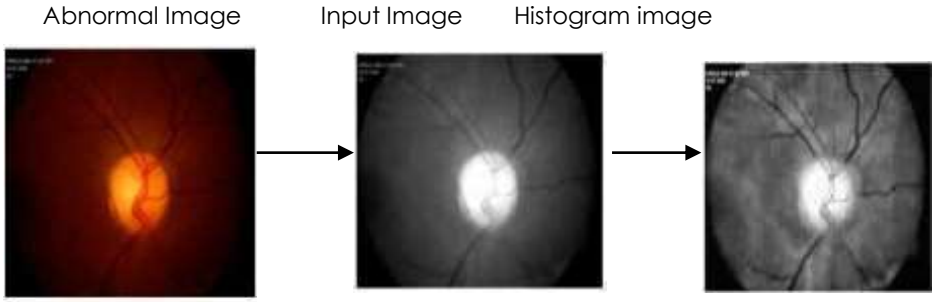


Fig.10: Pre-processed output for abnormal image

Feature Extraction

Feature extraction is a special form of dimensionality reduction. The purpose of feature extraction is to reduce original data set by measuring certain features that distinguish one region of interest from another. Wavelet based textural features namely energy, entropy, skewness, kurtosis are extracted from retinal images. The analysis and characterization of textures present in the medical images can be done by using the combination of Wavelet Statistical Texture features (WST) obtained from 2-level Discrete Wavelet Transformed (DWT) low and high frequency sub bands. A feature extraction is the determination of a feature or a feature vector from a pattern vector. In order to make pattern processing problems solvable one needs to convert patterns into features, which become condensed representations of patterns, ideally containing only salient information.

Wavelet Based Feature Extraction

A wavelet is a wave-like oscillation that is localized in the sense that it grows from zero, reaches maximum amplitude, and then decreases back to zero amplitude again. It thus has a location where it maximizes, a characteristic oscillation period, and also a scale over which it amplifies and declines. Wavelet analysis developed in the largely mathematical literature in the 1980's and began to be used commonly in geophysics in the 1990's. Wavelets can be used in signal analysis, image processing and data compression. They are useful for sorting out scale information, while still maintaining some degree of time or space locality. Wavelets can be used to compress the information in two-dimensional images from satellites or ground based remote sensing techniques such as radars.

Discrete Wavelet Transform

Discrete wavelet transform (DWT), which transforms a discrete time signal to a discrete wavelet representation, it converts an input series x0, x1,xm, into one high-pass wavelet coefficient series and one low-pass wavelet coefficient series (of length n/2 each) given by:

Where sm (Z) and tm (Z) are called wavelet filters, K is the length of the filter, and i=0,[n/2]-1. Lifting schema of DWT has been recognized as a faster approach.The basic principle is to factorize the poly phase matrix of a wavelet filter into a sequence of alternating upper and lower triangular matrices and a diagonal matrix. This leads to the wavelet implementation by means of banded-matrix multiplications.

Decomposition

Decomposing an image into meaningful components is an important and challenging inverse problem in image processing. A first range of models are denoising models: in such models, the image is assumed to have been corrupted by noise, and the processing purpose is to remove the noise. The decomposition process is mainly used to splitting or segmenting the given images.

Trace Transform

The TT can be defined as a functioning based on T, which is some functional of the image with variable t. T is called the trace functional. In order to define a triple feature, two more functional have been defined and they are designated by P .This is known as the diametrical functional, which is a functional of the TT function when it is considered as a function of the length of the normal to the line only called the circus functional, is a functional operating on the orientation variable, after the previous two operations (T and P) have been performed.

Higher Order Spectra

HOS is a nonlinear method which captures subtle changes in image pixels. The algorithm discussion starts with second order statistics which evaluate both mean value (m) and variance (σ2).

$$Ma = E\{A\} \quad (3)$$

$$\sigma 2A = E\{(A- Ma)2\} \quad (4)$$

In addition to these moments, HOS provides higher order moments, i.e., m3;
m4; . . . and nonlinear combinations of the higher order moments called “cumulants",i.e., c1; c2; c3; . . .

Thus, HOS consists of both moment and cumulant spectra.The technique can be used for deterministic and random signals. The so-called “bispectrum", which is a third order statistic, was used in this work. It is obtained by calculating the Fourier transform of the third order correlation of the data:

$$B(f1,f2) = E\{A(f1)\ A(f2)\ A^*(f1+ f2)\} \qquad (5)$$

where A(f) is the Fourier is transform of the signal a(nT) and E{.} is an average over an ensemble of random signal realizations. For deterministic signals, the relationship holds without an expectation operation. In this case, the third order correlation is a time-average. For deterministic sampled signals, A(f) is the discrete-time Fourier transform, which, in practice, is computed using the fast Fourier transform (FFT) algorithm. The frequency (f) may be normalized by the Nyquist frequency to be between 0 and 1. In this work, we derived the bispectral phase entropy (Ph), entropy 1 (P1), entropy2 (P2), and entropy 3 (P3). These entropies are similar to the spectral entropy.The equations which govern the phase entropy extraction from HOS parameters.

Scope

In existing method HOS,TT and DWT with SVM classifier only produce a 90% accuracy,87%sensitivity and 90% specificity were achieved. But in this proposed expectation Maximization classifier may produce 95% accuracy,92% sensitivity and 97% specificity.

RESULTS

Fig.11: Original Image

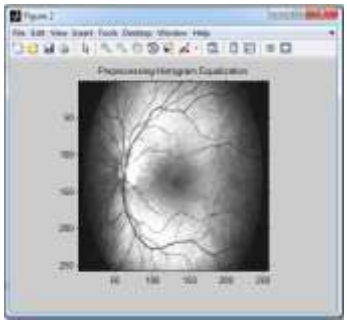


Fig.12: Histogram Equalization

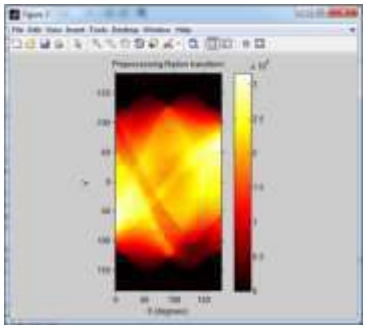


Fig.13: Radon preprocessing Image

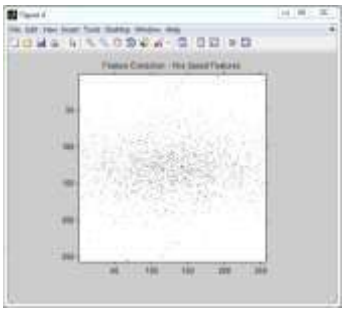


FIG.14: HOS Based Features Image

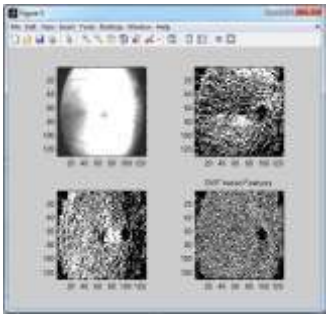


FIG.15: DWT Based Features Image

Take the original image from the fundus camera. Here focussing the retinal area of the eye. Histogram equalization focussed the intensity level of original image. Radon transform also pre-processed in varies degrees of retinal images.HOS is a nonlinear method to captures subtle changes in image pixels shown in fig11- 15. Based on the pixel value calculate the entrophy1, 2, 3 derived. A DWT show varies aspects of features like skewness, kurotis, symlet, and coiflets.

CONCLUSION

An automated system has been successfully developed which is able to detect the glaucoma from the retinal photographs with the performance approaching that of trained clinical observers. It has been found that the glaucoma can be detected irrespective of the stages of its growth. The DWT, TT, HOS based features been employed to detect the complication caused due to glaucoma. This method is found to reduce the manual effort required for the detection and also increase the accuracy when compared to previous method. The performance of the fully automatic system presented here is comparable to medical experts in detecting glaucomatous eyes and it could be used in mass-screenings. The important features automatically identified by the methods also provide a novel representation of the data for the physicians

and may help to better understand glaucoma. The detection and classification of glaucoma will done by DWT, TT, HOS transformations. The hybrid feature selection provides a powerful detection and diagnoses the diseases from the retinal images. The extracted hybrid features are fed to Expectation Maximization classifier to find the normal/glaucoma image with high accuracy, sensitivity and specificity. This system can be implemented in hospital to reduce the visual loss by helping the human environment with low cost of finding the glaucoma.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None.

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] AntoBennet M, Sankar Babu G, Natarajan S. [2015] Reverse Room Techniques for Irreversible Data Hiding", *Journal of Chemical and Pharmaceutical Sciences*. 08(03): 469-475.

[2] AntoBennet M, Sankaranarayanan S, Sankar Babu G. [2015] "Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis", *Journal of Chemical and Pharmaceutical Sciences*. 08(03): 571-576.

[3] AntoBennet M, Suresh R, Mohamed Sulaiman S. [2015] "Performance &analysis of automated removal of head movement artifacts in EEG using brain computer interface", *Journal of Chemical and Pharmaceutical Research*. 07(08): 291-299.

[4] AntoBennet M & JacobRaglend. [2012] Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", *Journal of Computer Science*. 8(9): 1447-1454.

[5] AntoBennet M & JacobRaglend.[2011.] Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", European Journal of Scientific Research, vol. 66 no.1, pp.120-129,

[6] AntoBennet M, Resmi R. Nair, Mahalakshmi V, Janakiraman G. [2016] "Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography", *Indian Journal of Science and Technology*. 09(02):01-07

[7] AntoBennet M & JacobRaglend. [2012] 'A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach', *Journal of Applied Sciences Research*. 8(5):2429-2438.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

NEW TECHNIQUE OF ACADEMIC DATA ANALYSIS BY FUZZY VARIANCE

Anil Kumar Tiwari[1], G Ramakrishna[2], Lokesh Kumar Sharma[3], Sunil Kumar Kashyap[4*]

[1,2]Department of Computer Science and Engineering, K L University, Vijayawada, INDIA
[3]National Institute of Occupational Health, Ahmadabad, Gujarat, INDIA
[4]Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology University, Vellore, Tamil Nadu, INDIA

ABSTRACT

Background: *The importance of academic data counts by the modern society. The analysis of the academic data is more important than previous statement. The objective of academic data analysis is to explore more information for future academic improvements. This paper deals with the fuzzy variance technique for the academic data analysis. The additional advantage of this technique is that it gives linguistic decision. Fuzzy variance interacts with data as the set and cluster as the subset. The subset say cluster of the set say data is optimized by the fuzzy variance and this is used for analyzing the performance of academic data of the students.* ***Methods:*** *Clustering the data is the mathematical method but its analysis is referred as the mining of data. The fuzzy variance is that mathematical method which enables the study into linguistic form also.* ***Results:*** *Thus, this paper proposes a new Data Mining Clustering Technique over the crisp data set of student's academic performances and then its interpretation by MATLAB. Its analysis over the fuzzy set is also presented.* ***Conclusions:*** *Hence, the optimized decision algorithm over the variance of Arithmetic-Fuzzy Variance is illustrated for the linguistic and numeric outcomes.*

KEY WORDS
Data mining technique, Crisp set, Fuzzy set, Optimization, Arithmetic fuzzy

Received: 11 Nov 2016
Accepted: 9 Feb 2017
Published: 8 March 2017

***Corresponding Author**
Email: 7sunilkumarkashyap@gmail.com
Tel.: +91-94242-16777

INTRODUCTION

The information can be represented into various discrete forms e.g. text, number, codes etc. This is said to be a data if it maps with the certain information. Data is the source of infinite information. Data is the set with an element minimum and n number of variables maximum. The element of the data referred as the co-domain of the function where the domain is the certain objective and range is another data. Every data generates the new data if there be applied a particular rule. Data is the storage of information. The redefining the sequence of the information of the given data is the challenge. To search a map from the single information to multiple information is the major objective of data mining. The study of data refers data mining. Data mining is the process to define the new information from the current information. It is science and art also. The set of rules referred as science and the sequence of the elements of that set referred as art. The Several application has been existing since 1979, when Hartign et.al. [1] designed an algorithm. It was the first K-means clustering algorithm. After a decade this is presented another clustering overview in the platform of Artificial Neural Network. It is analyzed in the perspective of statistics. In 1993, Jang [2] proposed an adaptive network based fuzzy inference system in the domain of physical environment. Gau et.al. [3] discovered a new set theory as Vague Sets in the same year. Law [4] used fuzzy numbers for grading the educational system, in the year 1996. In the same year, Possey et.al. [5] drafted a student model, which is based on the Neural Network.

2006 become the year of real application of K-means clustering in students academic performances. It was a ase study of pharmacy students. Its result and affect gave by Sansgiry [6]. Oyelade et.al. [7] presented an application of K-means clustering algorithm for prediction of the performances of the students, in the year 2010. Next year, Mankad et.al. [8] displayed an educational case study based on the genetic fuzzy algorithm. Next year, an improved academic performance system came in the existence. Defence University's data warehousing and data mining were the key operators of this system. In the same year, Choudhary et.al. developed a model based on soft computing for academic performances of teachers. It is also based on fuzzy logic. Upadhyaya [9] whown a result on fuzzy logic based evaluation of student's academic performances.

In this paper, an optimized fuzzy K-means clustering algorithm presents for the evaluation of student's academic performances. We analysed K-means, Fuzzy and Optimization Theory in the context to each other and classified the common structure.

The Intelligence is one of the qualities of any student, which becomes the key factor in the academic performances. Intelligence is one of the class or cluster. If we classify any class, then the intelligence plays an important role for grouping the class. It order is also an important phase of study. It can be studied as, either poor, average, intelligence or Intelligence, average, poor or, average, poor, intelligence or poor, intelligence, average or average, intelligence, poor or intelligence, poor, average. These can be formed as the elements of set but its conclusion always distinct with each other. For example, intelligence means score lies in the interval of 80-100, average means. Score lies in the interval of 60-80 and 40-60 for the poor. Suppose a student scores 40 marks in the quarterly examination, 60 marks in the half yearly examination and 90 marks in annual examination. Our conclusion will be as "his academic performance is improving". In other example, a student scored 90, 60, and 40 in the quarterly, half yearly and annual examination respectively. Then, definitely the conclusion will not as earlier. For this, we can say, the academic performance of student is reducing. The rate of learning is distinct in both the cases. Hence, one

objective is determining that the class homogeneity and academic performances can be functioned. The involvement of the linguistic variable, fuzzy logic interacts with this by the common application.

Finally, this paper process a new structure of fuzzy K-means and its application is presented as the algorithm for the efficient evaluation of student's academic performances. Mathematically, the presentation is based on the transformation from hypothetical to physical form as, let a set of n-data points in the space k is any integer. Then the objective is defined as "obtain k-points set in ". It calls centre point and hence, minimize the distance from each point to its nearest.

MATERIALS AND METHODS

There are various method for the similar objectives. It can be classified into quantitative and qualitative. Our objective is to find the decision in linguistic and numerical both, thus the following existed method is essential to read:

K-means clustering method

In section 2, it is defined, this section, it will be presented in operational way.
Let X be set and $\{A_i\}$, where $i = 1,2,...,C$ be a family set. Its partition will be represented as,

$$\bigcup_{i=1}^{C} A_i = X,$$
$$A_i \cap A_j = \phi; \forall i \neq j,$$
$$\phi \subset A_i \subset X; \forall i,$$

Where, $X = \{x_1,...,x_n\}$ is a data sample and C is the number of clusters.
For C; $2 \leq C < n; C = n.$

Hence, the objective function be,

$$J(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{C} \chi_{ik} (d_{ik})^2 \ ,$$

Where, U is the partition matrix, v is a vector of cluster centre and d_{ik} a Euclidean distance measure between x_k and v_i, it is represented as,

$$d_{ik} = d(x_k - v_i) = \|x_k - v_i\| = \sqrt{\sum_{j=1}^{m} (x_{kj} - v_{ij})}$$

Fuzzy C-means clustering method

Its fuzzy based C-means clustering method for the multiple clusters. Its objective function will be,

$$J(U,V) = \sum_{i=1}^{k}\sum_{x_k \in X} (\mu_{C_i}(x_k))^m \|x_k - v_i\|^2$$

Where, U is a fuzzy position and m is a weight.
For the local minimum, the objective function becomes;

$$\mu_{C_i}(x) = \frac{1}{\sum_{j=1}^{k} \left(\frac{\|x - v_i\|^2}{\|x - v_j\|^2} \right)^{\frac{1}{m-1}}}; 1 \leq i \leq k, x \in X$$

$$r_i = \frac{\sum_{x \in X} \left(\mu_{C_i}(x)\right)^m x}{\sum_{x \in X}^{n} \left(\mu_{C_i}(x)\right)^m}; 1 \leq i \leq k,$$

$$\sum_{i=1}^{C} \left\| r_i^{previous} - v_i \right\| \leq \varepsilon.$$

Fuzzy mean

As arithmetic mean, fuzzy arithmetic mean is the generalized form the first. Let a multiple be $(x, \mu(x), [0,1], l_1, l_2)$, where l_1, l_2 are the two linguistic variables. Then the Fuzzy Arithmetic Mean can be represented as consequence matrix in below::

$$\bar{f} = \begin{bmatrix} f_1 & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & f_n \end{bmatrix},$$

Where $f_1, \ldots f_n$ are the grading.

Trace of the fuzzy mean is equal to the sum of the diagonal element, i.e. $Tr(\bar{f})$.

Below, we propose a new method, which deals with the decision of numerical and linguistic both. Its explained by the sample data of a student, but it can be generalized for n variables.

RESULT

Proposed method

Table 1: An observation of a student's academic performance is presented

Test	I	II	III	IV
Marks (100)	20	50	80	?

There are the following natural questions:

a. What will be the marks in Test IV?
b. The average of marks is constant (50) from Test I to III and from III to I also, but what about the performance of student is improving or reducing?

This case study is mentioned earlier in this paper, but in this section, it will be analyzed. For searching the answer of these questions, the following methodology is essential:

Table 2: The classical set theory and modern fuzzy logic applied on the above observation as below

Test	I	II	III	IV
Marks (100)	20	50	80	?
Crisp Set	0	0	1	?
Fuzzy Set	0	0.5	1	?

The objective of given the above table is, "for identify the intelligent student". If the student scored more than or equal to 80 marks then it referred as "an intelligent student", thus crisp and fuzzy are existed here. Let a data set (Crisp Set) of the marks of student be $X = \{x_1, \ldots, x_n\}$ and another data set (Fuzzy Set) of the marks of student be $F = \{(x_1, \mu_A(x_1)), \ldots, (x_n, \mu_A(x_n))\}$. The arithmetic mean of the crisp set and fuzzy set be $\bar{x}$ and $\bar{f}$ respectively. The simulated marks of the student be x_m. It can be presented mathematically as,

$$x_m = f(\bar{x} \pm a); a \in R.$$
$$either, x_m = f(\bar{x} + a)$$
$$or, x_m = f(\bar{x} - a).$$

Similarly for the fuzzy set,

$$x_m = f(\bar{f} \pm a); a \in R.$$
$$either, x_m = f(\bar{f} + a)$$
$$or, x_m = f(\bar{f} - a).$$

A mapping a is defined from $\bar{x}$ to $\bar{f}$ as, $a : \bar{x} \to \bar{f}, or, \bar{f} = a(\bar{x})$ and it variance be v_x, v_f. Hence, x_m can be optimized as,

$$x_m = \max(v_x \cup v_f, \mu_{x_m}(v_x \cup v_f)),$$
$$s.t.,$$
$$v_x = \frac{\sum (d\bar{\bar{x}})^2}{N}; Population$$
$$Or,$$
$$v_x = \frac{\sum (d\bar{\bar{x}})^2}{N-1}; Sample$$
$$\&$$
$$v_f = \frac{\sum (d\bar{\bar{f}})^2}{N}; Population$$
$$Or,$$
$$v_f = \frac{\sum (d\bar{\bar{f}})^2}{N-1}; Sample$$
$$\&$$
$$Error : either, v_x, v_f \to <,$$
$$or, v_x, v_f \to > .$$

Its algorithm is presented in below section.

Algorithm

Input Set X : {(Student's Name, Marks)} = $\{(s_1, t_1), ..., (s_n, t_n)\}$.

Input Set F : {(Student's Name, Membership Function of Marks)} = $\{(s_1, \mu(t_1)), ..., (s_n, \mu(t_n))\}$.

Compute $\bar{x}, \bar{f}$.

Optimized by,

$$x_m = \max(v_x \cup v_f, \mu_{x_m}(v_x \cup v_f)),$$
$$s.t.,$$
$$v_x = \frac{\sum (d\bar{x}_n)^2}{N}; Population$$
$$Or,$$
$$v_x = \frac{\sum (d\bar{x}_n)^2}{N-1}; Sample$$
$$\&$$
$$v_f = \frac{\sum (\mu d\bar{x}_n)^2}{N}; Population$$
$$Or,$$
$$v_f = \frac{\sum (\mu d\bar{x}_n)^2}{N-1}; Sample$$
$$\&$$
$$Error : |ND - LD| \neq 0.$$

Where, ND and LD for Numeric and Linguistic Decisions respectively.

Output:

Numeric Decision:

$$x_m = f(\bar{x} \pm a); a \in R.$$
$$either, x_m = f(\bar{x} + a)$$
$$or, x_m = f(\bar{x} - a).$$

Linguistic Decision:

$$x_m = f(\bar{f} \pm a); a \in R.$$
$$either, x_m = f(\bar{f} + a)$$
$$or, x_m = f(\bar{f} - a).$$

Illustration is given below, which interacts with the above mentioned example.
Example:

Input Crisp Set X :

Input Fuzzy Set F :

CONCLUSION

Individually, statistical means used as the common tool in general decision theory. Fuzzy mean is the newly launched method. Its application is presented in this paper for the better decision than existed. The study of the academic performance of students is equivalent to the study of discrete function of arithmetic and fuzzy means. Both is taken simultaneously is still challenging, but the paper is become the first step toward this under the error estimation.

CONFLICT OF INTEREST
None

ACKNOWLEDGEMENTS
To the Laboratory, KL University, Vijayawada, Andhra Pradesh, INDIA.

FINANCIAL DISCLOSURE
This is an unfunded research.

REFERENCES

[1] Hartigan JA, Wong MA. [1979] A K-means clustering algorithm, Appl Stat, 28: 100-108.

[2] Jang JSR. [1993] ANFIS: Adaptive-network based fuzzy inference system, IEEE Trans Sys, Man, Cybernetics, 23(3): 665-685.

[3] Gau WL, Buehrer DJ. [1993] Vague sets, IEEE Transactions System, Man Cybernetics, 23(2): 610-614.

[4] Law CK. [1996] Using fuzzy numbers in educational grading system, Fuzzy Sets Syst, 83: 311-323.

[5] Posey CL, Hawkes LW. [1996] Neural networks applied in the student model, Intell Sys 88: 275-298.

[6] Sansgiry SS, Bhosle M, Sail K. [2006] Factors that affect academic performances among pharmacy students, Am J Phar Edu, 70(5): 231-243.

[7] Oyelade OJ, oladipupo OO and Obagbua IC. [2010] Application of K-means clustering algorithm for prediction of students's academic performances, Int J Comp Sci Inf Sec, 7(1): 292-295.

[8] Mankad K, Sajja PS, Akerkar R. [2008] Evolving rules using genetic algorithm: a hard problem for partition based approach, Int J Soft Comp Appl, 2: 6-15.

[9] Upadhyaya MS. [2012] Fuzzy logic based on performances of students in college, J Comput Appl 5(1): 6-9.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

NEW ACADEMIC PREDICTION SYSTEM BASED ON COMPACT SOFT COMPUTING

Swati Jain[1], Vikas Kumar Jain[2], Sunil Kumar Kashyap[3]*

[1]*Department of Computer Science, Kalinga University, Raipur, Chhattisgarh, INDIA*

[2]*Department of Chemistry, Government Engineering College, Raipur, Chhattisgarh, INDIA*

[3]*Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology University, Vellore, Tamil Nadu, INDIA*

ABSTRACT

Background: *The prediction of data is a science. Data is studied in this paper under the compact soft computing. The compact soft computing involves the fuzzy logic, fuzzy set, neural network and genetic algorithm. This paper delivers a system for predicting and evaluating the academic performance of the student. Hence the optimized data interpretation system is proposed in this paper.* ***Method:*** *This academic prediction system is based on compact soft computing.* ***Results:*** *A new academic prediction system is proposed in this paper which is based on the data analysis via compact soft computing.* ***Conclusions:*** *The fuzzy logic, fuzzy set, neural network and genetic algorithm are used as the compact computation tool in this paper. This is proved as the errorless interpretation.*

KEY WORDS
Compact Soft Computing, Fuzzy, Neural Network, Genetic Algorithm

Received: 11 Nov 2016
Accepted: 9 Feb 2017
Published: 8 March 2017

***Corresponding Author**
Email: 7sunilkumarkashyap@gmail.com
Tel.: +91-94242-16777

INTRODUCTION

The objective of the data analysis is to transform the data into useful and usable form. This is a transformation from fact to information. No data able to speak about themselves. Thus its interpretation is required. May be many information involved in single information or single information be a source of several information. These all hypothetical remarks studies in data analysis. Chow et al [1] studied the data under the probability distribution over the tree in 1968. This presented an approximation of the data and error estimation. Nakhaeizadeh [2] developed a data management system for banking in 1998. The data of banks inter-relate and co-relate by this data management system. The economical application of database first time came in the existence.

Nauck et al [3] established the neuro-fuzzy data base system in 1997. The fundamental rules were defined in this foundation. In 1988, Lauritzen et al [4] proposed the concept of local computation over the expert systems by probability. The graphical structures of the data and its application has presented in this concept. The previous Neruo-fuzzy concept used again for this, but this was presented as the survey of all such types of data system based results. In 1992, a noteworthy idea came in the existence as the Intelligent System. This system is invented by Pearl [5]. This is basically a reasoning based data base mechanism. In the same year, another mechanism came in the existence as learning system by examples. Wang et al [6] developed this theory through the fuzzy rule generator. The learning by example is the key idea behind this system.

Kruse [7] put the theory on data and its interpretation by the uncertainty and vagueness concept. The knowledge management system was launched as the mapping between uncertainty and vagueness in the year 1991 for the generation of the new database system. Three years later, they proposed another data management system based on fuzzy systems. Hackerman et al [8] generalized the data base which is recalled later as Bayesian Network. The combination of knowledge and statistical data is proposed first time as the learning management system in 1995. In 1999, Hoopner et al [9] presented the data analysis by the cluster algorithm. The cluster and fuzzy presented simultaneously in this survey.

The possibilistic model is developed as the second model. The data and network were the two sets for establishing the map. The hyper tree decomposition model was the second last model in the series. This model was based on multivariate possibility distribution. The last model was the source of information. By the parallel combination this was structured.

In 1999, Gentch [10] proposed some tools for data mining. In the same year, Dubois et al [11] merged the fuzzy information. Fayyad et al [12] edited a note on modern data management system in 1996. Anderson et al [13] proposed an expert system based on HUGIN. In 1989, they generalized the data over the Bayesian Distribution Process (BDP).

Next section deals with pre-requisites on soft computing techniques.

MATERIALS AND METHODS

The Analysis of Fuzzy Models: Fuzzy set, Neural Network and Genetic Algorithm are the base of this paper. Its composition for data mining of student's academic performance is the objective of this paper.

The sequence is remaining as per the mentioned, first Fuzzy Set, then Neural Network and Genetic Algorithm in the last. It is defined in below:

Fuzzy Set: Let X be a space of points, with a generic element of X denoted by x. Thus, $X = \{x\}$.

A fuzzy set A in X is characterized by a membership function $f_A(x)$ which associates with each point in X a real number in the interval [0, 1] with the value of $f_A(x)$ at x representing the grade of membership of x in A.

Fuzzy Logic: According to Zadeh [31], Fuzzy Logic is a logical system for forming the system by approximate reasoning.

Fuzzy is method of redefining the class according to the membership. The class and the membership is decided by the social and understanding behavior. The classical set never behaves as variable but the fuzzy set always holds the same always by the logic. In general, the crisp value, e.g. 2 means only 2 again in the classical set theory but 2 means 2, 2.1, 2.2, 2.3, 2.4, 2.7, ...2.9, ...2.11...2.99 and so on...in fuzzy theory.

Neural Network: It is inspired by the mechanism of human brain. The layers are the elements of any NN. It is also referred as Artificial NN, by the reason of it is not natural brain but an artificial.

Basically the followings are the main steps of ANN:

1. Input Layer
2. Middle Layer
3. Output Layer

By the structure it can be demonstrated as follows:

The Activation Function: Let the input variables be x_i and its corresponding weight be w_i, then the weighted sum is called activation function and it is represented by,

$$A(\bar{x}_i, \bar{w}_i) = \sum_{i=0}^{n} x_i w_i .$$

The Sigmoidal Function: An activation function is called a Sigmoidal Function if it is represented as,

$$O(\bar{x}_i, \bar{w}_i) = \frac{1}{1 + e^{A(\bar{x}_i, \bar{w}_i)}}$$

The Error Function: The sum of all the layers of output is called an error function if it represented as,

Genetic Algorithm: Fitness proportionate selection, recombination/crossover and mutation are the three fundamental characteristics of GA.

$$E(\bar{x}_i, \bar{w}_i, d) = (O(\bar{x}_i, \bar{w}_i) - d)^2$$

Definition: Let a function $f : X \to R \geq 0$ be given, then the optimization problem is represented by,

$Optimize(x) = \arg(\max(f(x)))$, Where $f(x)$ is the fitness function.

Definition (Proportionate Selection): The rate of probability is represented by,

$$. \; p(x, t+1) = p(x,t) \frac{f(x)}{f(t)}$$

Definition (Proportionate Selection Function): The response function for the proportionate function is represented by,

$$R(t) = \frac{V(t)}{\bar{f}(t)},$$
$$V(t) = \sum_x p(x,t)(f(x) - \bar{f}(t))^2$$
$$R(t) = \bar{f}(t+1) - \bar{f}(t)$$
$$\bar{f}(t) = \sum_x p(x,t) f(x)$$

Definition (Recombination): Robbin's distribution $\pi(x,t)$ is given by,

$$\pi(x,t) = \prod_{i=1}^{n} p_i(x_i,t).$$

The generalization of the genetic algorithm as per the dynamic data analysis is presented in the below section:

Generalized Genetic Algorithm: It is the process for generating the best variable for the next operation. The data of student's academic performance is studied as the application of GA.

Let the random variables are: $x_1, x_2, x_3, \ldots x_n$.

The corresponding weights are: $w_1, w_2, w_3, \ldots w_n$.

The summation of the weights is:

$$\sum_{i=1}^{n} w_i = w_1 + w_2 + w_3 + \ldots + w_n$$

or,

$$\sum = w_1 + w_2 + w_3 + \ldots + w_n.$$

Hence, the Activation Function is: $f(\sum)$.

Next is the fitness function, which is required for executing the process.
Let S be a total number of samples, G be the global error, t_i be the time at i position. Hence the fitness function will be;

$$f = \frac{1}{G} = \frac{1}{\sum_{i=1}^{s} (t_i)^2}.$$

The crossover function is defined in the next section.

Let X_i, X_{i+1}^t be the pair before crossover, X_i^{t+1}, X_{i+1}^{t+1} be the pair after crossover, C_i be the random number of uniform distribution in [0, 1], then,

$$X_i^{t+1} = c_i . X_i^t + (1 - c_i) . X_{i+1}^t$$
$$X_{i+1}^{t+1} = (1 - c_i) . X_i^t + c_i . X_{i+1}^t.$$

The algorithm is given in the next section.

1. The given data.
2. The Set of Random Numbers.
3. The coding by real numbers by; $L = i \times s + s \times j,$

Where,

i be the Input random number, s be the sample random number and j be the out random number.

4. $\min(f)$.
5. New Population Generation by;

$$X_i^{t+1} = c_i . X_i^t + (1 - c_i) . X_{i+1}^t$$
$$X_{i+1}^{t+1} = (1 - c_i) . X_i^t + c_i . X_{i+1}^t.$$

Next section lies with the proposed system.

RESULTS

Data Set:

$$L = \begin{bmatrix} l_{11} & \cdot & \cdot & \cdot & l_{1m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ l_{n1} & \cdot & \cdot & \cdot & l_{mn} \end{bmatrix}_{m \times n}$$

Fuzzy Set:

$$A = \begin{bmatrix} \mu_{11} / x_{11} & \cdot & \cdot & \cdot & \mu_{1n} / x_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mu_{m1} / x_{m1} & \cdot & \cdot & \cdot & \mu_{mn} / x_{mn} \end{bmatrix}_{m \times n}$$

The Real Number Coding: $L = i \times s + s \times j,$

Where,
i be the Input random number, s be the sample random number and j be the out random number.

Optimization:

$$\min f = \frac{1}{L} = \frac{1}{\sum_{i=1}^{i} (x_i)^2}$$

New Population Generation:

$$X_i^{t+1} = c_i . X_i^t + (1 - c_i) . X_{i+1}^t$$
$$X_{i+1}^{t+1} = (1 - c_i) . X_i^t + c_i . X_{i+1}^t .$$

Predicted Data:

$$(\pi(x,t), L) = \left(\prod_{i=1}^{n} p_i(x_i, t), A \right)$$

.

CONCLUSION

The data is not just a fact but more than the fact. The past, present and future data can be mapped with each other. The probability on the data is applied not only for studying the certainty and uncertainty of data but to define the data as the real value. The prediction of the data is important because repetition could be avoided. It is all an optimization. The data then its classification by fuzzy then its weight consideration and then the best fit data or the origin of the data or the generator of the data or the mean data or the inference data or the central valued data or an only data, which could be presented or leaded as the universe of the data. Thus compact model fulfils the desired goal.

CONFLICT OF INTEREST
None.

ACKNOWLEDGEMENTS
Thanks to the reviewers for their valuable suggestions.

FINANCIAL DISCLOSURE
This is an unfunded research.

REFERENCES

[1] Chow CK, Liu CN. [1968] Approximating Discrete Probability Distributions with Dependence Trees. IEEE Trans. on Information Theory. IEEE Press, Piscataway, NJ, USA. 14(3) 462–467.
[2] [1998] Nakhaeizadeh. Wissensentdeckung in Datenbanken und Data Mining: Ein Uberblick. In: G. Nakhaeizadeh, ed. Data Mining: Theoretische Aspekte und Anwendungen, Physica-Verlag, Heidelberg, Germany. 1–33.
[3] Nauck D, Kruse R. [1998] Chapter D.2: Neuro-fuzzy Systems. In: E. Ruspini, P. Bonissone, and W. Pedrycz, eds. Handbook of Fuzzy Computation. Institute of Physics Publishing Ltd., Philadelphia, PA, USA.
[4] Lauritzen SL, Spiegelhalter DJ. [1988] Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Stat. Soc., Series B Blackwell, Oxford, United Kingdom. 2(50):157–224.
[5] Pearl J. [1992] Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition). Morgan Kaufmann, San Mateo, CA, USA.
[6] Wang LX, Mendel JM. [1992] Generating Fuzzy Rules by Learning from Examples. IEEE Trans. on Systems, Man, and Cybernetics, IEEE Press, Piscataway, NJ, USA. 22(6):1414–1427.
[7] Kruse R, Borgelt C, Nauck D. [1999] Fuzzy Data Analysis: Challenges and Perspectives. Proc. 8th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'99, Seoul, Korea). IEEE Press, Piscataway, NJ, USA (to appear).
[8] Heckerman D, Geiger D, Chickering DM. [1995] Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning, Kluwer, Dordrecht, Netherlands. 20:197–243.
[9] Hoppner F, Klawonn F, Kruse R, Runkler T. [1999] Fuzzy Cluster Analysis. J. Wiley & Sons, Chichester, United Kingdom.
[10] Gebhardt J, Kruse R. [1993] The Context Model — An Integrating View of Vagueness and Uncertainty. Int. Journal of Approximate Reasoning. North-Hollan, Amsterdam, Netherlands. 9:283–314.
[11] Dubois D, Prade H, Yager RR. [1996] Information Engineering and Fuzzy Logic. Proc. 5th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'96, New Orleans, LA, USA), IEEE Press, Piscataway, NJ, USA. 1525–1531.
[12] Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, eds. [1996] Advances in Knowledge Discovery and Data Mining. MIT Press, Menlo Park, CA, USA.
[13] Andersen SK, Olesen KG, Jensen FV, Jensen F. [1989] HUGIN — A Shell for Building Bayesian Belief Universes for Expert Systems. Proc. 11th Int. J. Conf. on Artificial Intelligence (IJCAI'89, Detroit, MI, USA), Morgan Kaufman, San Mateo, CA, USA. 1080–1085.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

THE ACTIVE IMPACT OF HUMAN COMPUTER INTERACTION (HCI) ON ECONOMIC, CULTURAL AND SOCIAL LIFE

Arun Kumar Singh
Asst. Professor, College of Computing and Informatics,
Saudi Electronics University, Abha City, KINGDOM OF SAUDI ARABIA (KSA)

ABSTRACT

Human Computer Interaction (HCI) is the technical feature of how human can interact or interface with their computing devices. Developer and programmers always have to search about HCI when developing new software application and products because inventing a new piece of equipment that is revolutionary and it becomes use to. Due to this effort and development of their products become user friendly and people have to consider HCI as easy to use by everyday life. One of the main feature within HCI is that how humans can use computers. Studies and research activities are under consideration to explore how people use and interact with computers and how new development would affect user's future communication and interaction. "Human Computer Interaction is well-organized with the design, evaluation and implementation of interactive computing devices for human use. Human Computer Interaction (HCI) research is performed to provide and promote a scientific understanding of the interaction between humans and the Computing/Informatics technology and tools that we use. HCI is an important and effective area to concern in human lives. Now a day it is necessary to be able to communicate information to each other, whether by reading text online, seeing videos or by using more advanced computer technology. In this article, I will try to describe the active impact of HCI on Economic, Cultural Environment and Social Networking life.

KEY WORDS
Human Computer Interaction (HCI), Social Life, Economic, Culture, Affect

Received: 22 May 2017
Accepted: 14 June 2017
Published: 30June 2017

***Corresponding Author**
Email:
a.singh@seu.edu.sa,
arunsinghiiita@gmail.com
Tel.: +966-593565533,
+91-7895276828
Fax: +966-172 417169

INTRODUCTION

"Human Computer Interaction is a well-organized with the design, evaluation and implementation of interactive computing devices for human use. Human Computer Interaction (HCI) research is performed to provide and promote a scientific understanding of the interaction between humans and the Computing/Informatics technology and tools that we use. HCI is an important and effective area to concern in human lives. Now a day it is necessary to be able to communicate information to each other, whether by reading text online, seeing videos or by using more advanced computer technology.
People have to consider HCI because the products concern with the easy interface and easy to use by peoples in their everyday life. [1] HCI consists mainly 3 parts:

1. The Customer
2. The Computing Devices
3. The Interactive Way that both Work Together

If all of these above concern parts are not coming together in a new software or product, then a perfect impact and success will be not happened in human computer interaction. HCI also has many fields these are where HCI works properly:

Psychology, Computer Science, Language, Sociology, Ethnography, Semiotics and Branding, Design, Engineering, Ergonomics and Human Factor. In Semiotics, instead of text this is the use of symbols, this is mainly used because people speak different languages and if symbols are used language isn't needed and there is less confusion [2]. Ethnography, this means that things in one country could mean something different in another country. An example of this is in the India it is common for an electric plug is Type C, D and M, while in Saudi Arabia it is common for an electric plug is Type A, B, F and G. So, it important in HCI to change specific things in different countries so the population isn't offended [3].

HCI impact giving a dramatic change on the way of people's economy, culture and their social life in this Computing/Informatics Era. In the case of virtual reality people believe that it will become the new reality which we live in, because it creates a new life for everyone even disabled users, they will now be able to live a normal life without any disabilities. HCI providing a positive effect on culture, economy and society, it has created new ideas and products making now a day for better and more advanced place to live in. We must accept that HCI will affect the future massively, in both a good way and in a bad way, for example new developed and improved types of product will be created that will take the society to a new era of technology and seem intelligent, but to create these types of products it will cost a lot of resources putting the planet in a bad way, it will cost lots of money to get the new products perfect and probably most of the planets resources emphasizing Global warming problems.

BASIC EXAMPLE OF THE HCI

One of the main focuses within HCI is the way humans' use computers and devices. Study development and research are undertaken to explore how human use and interact with computing devices and how new designs would affect user's future interfacing.

IIOAB JOURNAL
ISSN: 0976-3104

There are many different interfaces and developed designs that allow people to interact with a computer, some of the more common/well known designs are: [4]

Terminal (Mac)/Command line Editor (CLE-Window)

The terminal/CLE was the only way that people could interact with a computer before the GUI was created.

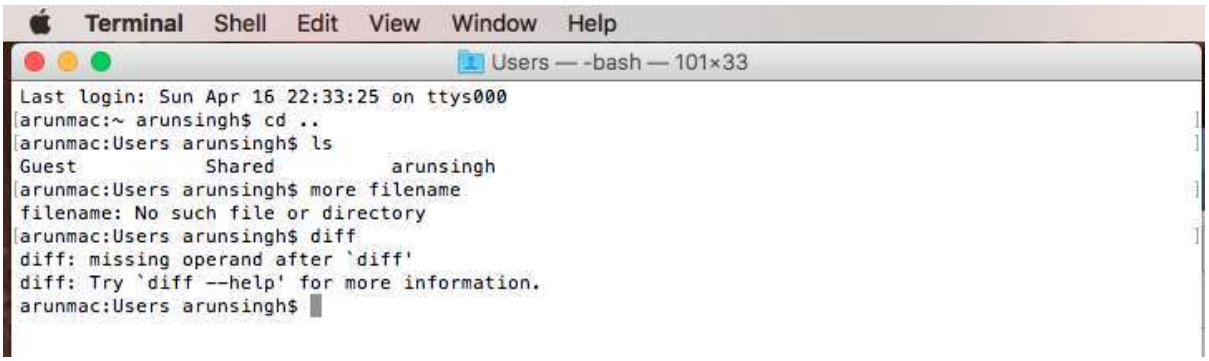


Fig.1: Terminal in MacBook Pro (Screen Shot 2017-04-16 at 11.02.46 PM).

The user would need to enter a series of commands into the operating system and press the Return key. The computer would then complete the required task; whether it's creating, copying, deleting or renaming files. As the terminal/CLE was text based, users had to input all of the commands without any typing errors, as a result this made the terminal/CLE an unfriendly way to interact with the computer.

Windows, Icons, Menus and Pointers (WIMP)

WIMP is used as a derogatory term to describe the graphical interface used by the Apple Macintosh and Microsoft Windows operating systems. The term WIMP was commonly adopted by users who strictly utilized text-based operating systems, such as a DOS, Linux, or Unix variant. [5]

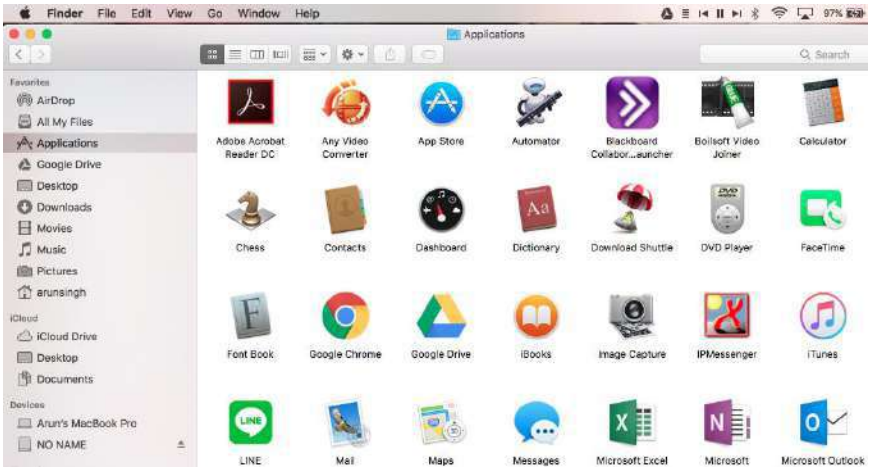


Fig. 2: WIMP in MacBook Pro (Screen Shot 2017-04-16 at 11.07.53 PM)

WIMP is a graphical way to use a computer, by using windows that display different Icons. The icon image changes depending on the file type or what program it will open when the icon is clicked on. Buttons are also represented by icons. Menus are also on the window, such as the 'File' menu, these give the user more options to do things like creating new folders. Everything within the windows can be selected or opened by clicking with the mouse (Pointer). When compared to CLE/Terminal interface, the WIMP is a lot more user friendly.

Graphical Users Interface (GUI)

The GUI makes use of graphics to aid the user when using the computer, these can be in the form of drop down menu's, ticking boxes or selecting/clicking on a button. GUI gives the user an easy way to access the features of a program, Microsoft Word relies on GUI a great deal for its toolbars and dialog boxes. GUI was designed to be used instead of the CLE as its easier to use and now almost all programs created uses a form of GUI. [6]

IIOAB JOURNAL
ISSN: 0976-3104

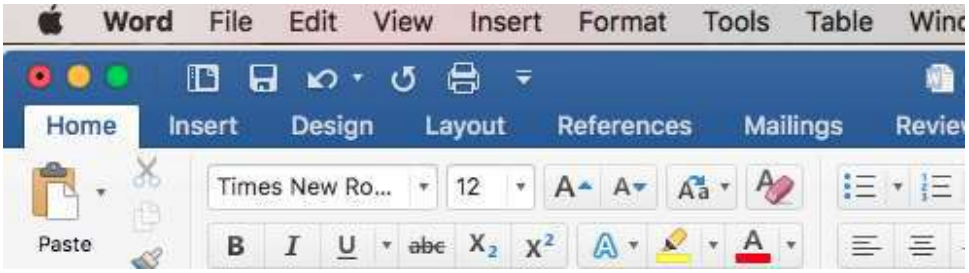


Fig. 3: GUI in Microsoft Office 365 (Screen Shot 2017-04-16 at 11.22.25 PM).

Web User Interface (WUI)

A good user interface, that is intuitive yet powerful, is essential when configuring and managing network security components. A browser based Web User Interface (WUI) represents a good way to provide this. Administrators and users already know how to use a web browser to access services. Providing network device configuration in this way builds on this experience and gives them a familiar, yet rich, way to configure and manage network resources.

Fig. 4: WUI in MacBook Pro (Screen Shot 2017-04-17 at 12.33.11 AM).

The user interface (UI), in the industrial design field of human computer interaction, is the space where interactions between humans and machines occur. The goal of this interaction is to allow effective operation and control of the machine from the human end, whilst the machine simultaneously feeds back information that aids the operators' decision-making process. Examples of this broad concept of user interfaces include the interactive aspects of computer operating systems, hand tools, heavy machinery operator controls, and process controls. The design considerations applicable when creating user interfaces are related to or involve such disciplines as ergonomics and psychology. [7]

ACTIVE IMPACT OF HCI

Economy

There are number of speed up inputs everywhere to help a user now. As an example of a speed up input we can see in a super market. This can help an individual because if ever one want to know the price of a product they can go to a barcode scanning machine where user can press the barcode to the reader and it tells about the detail of product. This impact on the economy will mean that people can shop much faster and efficiently and will encourage older people to shop as they will not have to Increased Automation of Outputs will mean that a business or company does not have to hire any employees or have to pay amounts of money to their employees as a company will only have limited employees to pay a salary to. [8] Decreased Complexity of Inputs will mean that people will be able to use new devices much more easily. It has to be kept simple otherwise new people that come will not know how to operate their devices and screen.

Text Readers impact the economy for visually and speech impaired individuals, a text read says aloud whatever you type into it. This will impact the economy because if a speech impaired person was to hold a meeting or a say a speech to a group of people they could easily understand the person due to the text

IIOAB JOURNAL
ISSN: 0976-3104

reader. Some new Text Readers now can translate something you type in English and say something in, for example Arabic or Hindi. This impact in the economy will help people interact to different people around the world and will hopefully help companies and business's grow.

Voice recognition or Voice Input is generally used in devices such a mobile to help send text messages and to communicate people and browse the internet web of concern contacts. This will affect that in buildings or famous places they can put voice recognition in elevators or in doors, so someone who is visually impaired can say "Open" and the doors will open up without pressing the buttons and for elevators a person could say "Floor 13" and it will take the person.

The impact of HCI on the economy will allow people with disabilities to do all activities everywhere and will help them to do jobs and other things. Mobile Communications allow people to communicate with others around the globe no matter day or night, the impact on the economy has helped companies to make communication which can help make and finalise deals of discussions. [9]

Culture

HCI has affected culture in many ways; Technology has opened up a lot of opportunities for private businesses because normally a small business would have too rent out an office to work, but nowadays they can do all of their work from home which is more convenient for them. People can now text each other to set up meetings instead of arranging one on the phone or by e-mail. This makes thing a lot easier to do because it is a faster and more efficient way of communicating. Because of technological developments, a greater number of people are able to work from home and are able to be more flexible and effective with where they use their computer. Laptops, PDA's, mobile phones, mobile computing and other portable devices, have replaced paper diaries and journals. Remote access to the Internet, home desktops and documents, such as word processing or spreadsheet files, have introduced people into mobile computing. As Internet access and speeds have increased, email has taken off which has mostly replaced letters and memos. [8, 9]

Laptops have becoming smaller, lighter, cheaper and more powerful over the years and has meant that they have become more common. They can be carried around with ease, this makes them an ideal for people who travel and still want to be able to access a computer or resources online. Not only are laptops being used for computing but also gaming and entertainment, particularly when traveling, but this has led to tablets (such as the iPad) which have started to replace laptops for portable gaming and entertainment. The way in which people are now buying and listening to music has changed, buying a physical album is no longer the main way which people listen to their music. The main way that albums are now being bought is online and downloaded onto the person's computer, one example would be iTunes by Apple. As well as the change to downloading music, portable music players (like the MP4 or iPod) have been developed and are becoming more common; as they allow the user to listen to their music when on the move. Games are becoming more complex and sophisticated and require more disc space to store the larger programs and require much more powerful hardware, in order to be played. Consoles where created into to run video games, these consoles allowed players to play games in their living rooms or any other room in their house. The current major released consoles are the Xbox 360 (Pro, S and E) by Microsoft and PlayStation 3/4 by Sony.

With the development of new HCIs', some working environments require less variety of work or skills. In some manufacturing companies use robots to construct cars (or other products that require large amount of assembling), as a result this has led to less specialized employees. This is known as deskilling. Automated systems can reduce the complexity of work, meaning the employees may feel that they are less valued and motivated.

Developing nations have seen big changes because of new technology. More call Centre's are based overseas and more products are being produced aboard and imported, this has led to an increase in job for those' living in developing counties.

Society

With the development of new HCI's, computers accessibility becomes more to people and so are being used more and more in everyday life. A large amount of the population own computers or have access to a computer to use for whatever purpose they need it for, whether that to word process documents, research material online, send emails. Computer chips are also used in a large range of electrical equipment for example domestic and industrial appliances from washing machines to microwaves to cook food where buttons and dials can be used as the interface. Many businesses, hospitals, researchers, creative industries (for example Movie, Games and TV companies) are just some of the people that depend upon computers. By creating new ways for people to interact with and could even be used in new ways.

But also, creating a new form of HCI, even if designed to be easier and more effective for human use, sometimes isn't picked up on quickly or sometimes avoided because it is designed to be used in a manner that some people find unusual or alien to them. That's why a lot of developers try to keep consistency in their HCI designs, so that people who have used one of their programs and then can open another one of

IIOAB JOURNAL
ISSN: 0976-3104

their programs and feel more comfortable, as the menus, options, colours, fonts and the overall interface look similar to their other program.

Even domestic appliances make use of simple and well-designed HCI's, as people normally understand how to use the appliance without even looking at the instruction manual before use. Not only do the domestic appliances use words to explain to the user what each button does but they can also use images, drawing or icons to inform the user about the function of each button. Using images can be more effective than using words, as images can be universally recognized, whereas words can only be understood by people who speak that language (for example the pause and play icons that have transferred from musical cassette players to computer software). Not only has the computer changed because of the developments in HCI, but almost everything that has a computer within it, from washing machines to microwaves to mobile phones. [8, 9]

Voice recognition software allows the user to speak, via a microphone to the computer in order to write text. The software takes what the user has said and then converts it into text on the screen, as if it's being typed by the user. Although this software can aid those with visual impairments there can be some drawbacks for the user, such as a delay between what has been said and the word being displayed on the screen or requiring the user 'train' the computer to accept their voice. Speech synthesis, similar to voice recognition, creates speech from word that the user has entered. An example of this would be the software that Stephen Hawking uses to interact with other people. The same method and artificial intelligence is also used within automated phone services, but instead of entering text the user is asked to press a different key in their phone. There are also interfaces that are designed that don't use voice recognition or speech synthesis, but make use of extra on-screen help. For example, subtitles can be turned on for TV programs movies and even some web sites, to help users that have hearing impairments. Some interfaces have been specially design for users with sight impairments by making use of larger text or using colours that are not affected by people who are colour blind. Red-green combination is the most common form of colour blindness, so by using a red-blue or yellow-blue colour combination in the interface design would allow a lot of people who are colour blind not to miss information and find using the interface easier. Remote controlled devices are often used for hostile environments that are too dangerous for humans, such as using bomb disposal robots to locate and dispose of explosives. Other forms of HCI can be used to allow for people to access inaccessible environments, like sending satellites into space or remote-controlled vehicle to Mars to conduct research. This means that the user can be a safe, in some cases a long distance away from the danger.

CONCLUSION

Human Computer Interaction (HCI) is the technical feature of how human can interact or interface with their computing devices. Developer and programmers always have to search about HCI when developing new software application and products, because inventing a new piece of equipment that is revolutionary and it becomes use to. Due to this effort and development of their products as user friendly, people have to consider HCI as easy to use by everyday life.

In the concern area of culture, effect of the fast inputs working on economy will make shopping faster in store, and much more efficient for the customers. This is making products purchasing easier. The HCI impact reduce the complexity of input on economy, if all products were made as user friendly and easy to access for everyone. The effect on a text reader will creates access to disabled users who cannot read text but need to hear it instead. The voice input software effect has on the economy is that it has created another type of market specified for the disabled user trying to make life as easy as possible for them. Mobile communication is providing a popular product for every user by making it easy communication. It becomes extreme popular in the world now a day. Playing games can help improve the capability and hand-eye coordination. They provide mental stimulation and promote social interactive networking. Flexible phone makes it easy for user to communicate and improved the usability that is the way of products that can made to be as easy to use as possible for every possible user. In other way touch screen effect have on society has become massive because it is easy to use, beneficial, and the latest product of this era. Input/output accessories has affected the society by making it easier to interact with mobile phones or other computing devices, if all products were not user friendly it would make it very difficult for people to use the product, this is why companies are manufacturing the standard size accessories for easiest way to use a product. Domestic computing appliances allow users to access around the house from a single place. Data logging is a right way for storing the data that may help researcher and developer later on; This will affect society by making it easier to log data. Virtual reality (VR) is also applies to computer-simulated environments that can it very simple to use, so that the user can get the full flesh effect of virtual reality.
"Human Computer Interaction is a well-organized with the design, evaluation and implementation of interactive computing devices for human use. Human Computer Interaction (HCI) research is performed to provide and promote a scientific understanding of the interaction between humans and the Computing/Informatics technology and tools that we use. HCI is an important and effective area to concern in human lives. Now a day it is necessary to be able to communicate information to each other, whether by reading text online, seeing videos or by using more advanced computer technology.

CONFLICT OF INTEREST
Authors confirm there is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
Authors confirm there are no financial disclosures.

REFERENCES

[1] (http://www.cadazz.com/cad-software-Sketchpad.htm) Information on Ivan Sutherland and the 'Sketchpad'. Cited 02/10/2012
[2] (http://www.guardian.co.uk/technology/gallery/2010/feb/14/mobilephones- gadgets-iphone) Information on mobile phones. Cited 02/10/2012
[3] http://asl.cs.depaul.edu/ (http://asl.cs.depaul.edu/) information on the hearing impaired, Cited 15/10/2012
[4] http://www.jhigh.co.uk/ComputingSG/GPPs/HCI/HCI_GUI.html
[5] http://en.wikipedia.org/wiki/Color_blindness#Design_implications_of_color_blindness
[6] http://www.baesystems.com
[7] https://en.wikipedia.org/wiki/User_interface
[8] https://rizyryan.files.wordpress.com/2012/10
[9] Karen Anderson, David Atkinson Beaumont, Allen Kaye, Jenny Lawson, Richard McGill, Jenny Phillips, Daniel Richardson. Information Technology Level 3 (Book 2). Heinemann (behalf of edexcel). ISBN 9781846909290

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

ANALYZING A PERSONALIZED NETWORK SYSTEM THROUGH NETFLOW

M. Jothish Kumar*, Baskaran Ramachandran

Department of Computer Science and Engineering, Anna University, Chennai, INDIA

ABSTRACT

Background: *In today's technological development Network has become crucial and also a ruling power. Monitoring and analyzing this network is cohesive. To overcome this kind of unmanageable task we need to undergo a process named netflow. Netflow is used to understand the difficulties we face through congestion in networks.* ***Methods:*** *Literally it helps to monitor the flow of network and collects detailed information being consumed by the users. In fact netflow gives an insightful view like who is the user, what kind of application is consumed, at which time, along with its source and destination i.e., IP address. Netflow helps to differentiate and peculiarize its user's consumption type, time and destination by averting congestion.* ***Results:*** *Through this personalized network system we can able to integrate CPU utilization, IP packet distribution, Protocol statistics, Top talkers and Protocol discovery.* ***Conclusions:*** *We can able to analyze and control the congestion in the network like bandwidth, throughput, packet loss and active flows. In fact netflow gives a insightful view of active users by averting congestion.*

KEY WORDS

Netflow,Congestion,CPU utilization,Bandwidth, Packet loss,Throughput

Received: 26 May 2017
Accepted: 18 June 2017
Published: 2 July 2017

***Corresponding Author**
Email: saijoetvm@gmail.com
Tel: 9787074363

INTRODUCTION

Today network congestion is crucial and tough to find its packet loss while data is transferred between the client and server. In companies, business people and workers expect a visibility network system-knowing its end user and understanding the flow of network without any traffic congestion. A high speed network with easy accessibility is always essential for a company's business. This paper states about a supporting system which offers a suitable solutions for the network issues.

The Netflow [1],[20] system posses application recognition with integrated trouble shooting features like scalability and extensibility to integrate other network congestion sources. Now-a-days rapid growth on network devices with increase in accessible performance makes an insane situation to internet service providers[2], due to network congestion. Especially piercing towards the spots and collecting data regarding network congestion becomes havoc. So by using the proposed netflow source we can easily identify the congestion and report about network operation.

MATERIALS AND METHODS

Netflow

Professionals feel netflow will be an indispensable tool on utilization of network resources. Netflow[3] can be defined as a stream of unidirectional packets between the given source and destination using IP address. But it is so critical to define how it works. While applying netflow we can easily analyze the long compliance issues and network anomaly. [Fig.1] shows that Netflow can be divided into three stages: Netflow Exporter, Netflow Collector, storage and segregate it to terminals based on their configuration of IP address.

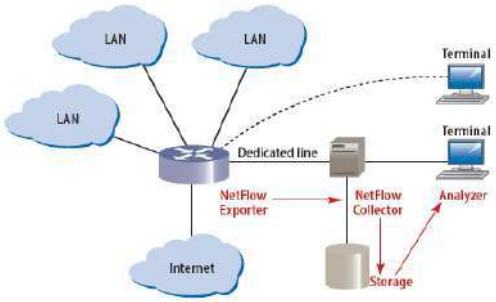


Fig. 1: NetFlow architecture. [4]

Netflow analysis

Netflow[5],[11] can be analyzed through a pair of performance by Network Management System (NMS). This system provides reports and alerts on trap viewers. Using a WAN link we can analyze the problem, when it happens, why and how it happened. It also gives alert when it is overloaded or fails to perform the

IIOAB JOURNAL
ISSN: 0976-3104

function. It will be received in the form of text messages or pages from NMS to NME (Network Management Engineer). From this message, we can able to link the user site with the headquarters and find out the reasons for the saturation of netflow.

Previously a special analyzer protocol will be linked with the WAN network. Later RMON analyzing protocol equipment is implant in the network system. RMON[6] is used only with LAN connections and WAN is used to create bandwidth issues frequently. Now netflow has commenced into this management system and plays a unique role in network system. This system is instigated by CISCO, a great sharing market in the field of network marketing.

NetFlow application and usage

In the field of technology netflow acts as a leading analyzer and accounter. It has numerous benefits based on its usage. It is used for security analysis, monitoring [7] the netflow ports, normalizing the packet tracks; reduce the data between source and destination in a single datagram record. NFC provides automatic metering and flexible key features to analyze and report about network operations. Netflow is a cost effective protocol used for dual purpose, i.e., switches and routes the packet tracks and produces netflow records. It helps to normalize between two end point IP addresses.

Netflow exporter

Netflow Exporter inspects the packets arrived and classifies it based on it attributes like source IP address, destination IP address , source port , destination port, Layer 3 protocol, class of service and input interface. After inspecting it is aggregated into flows and store in netflow cache as flow records. Then it is send to netflow collector

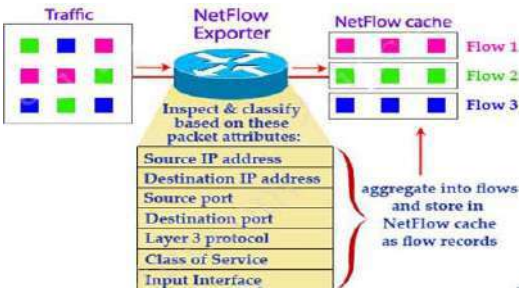


Fig. 2: Aggregate of NetFlow[8].

Netflow Cache

NetFlow cache shows how the traffic is utilized by the network device based on the source address, destination address, class of service to the priority of the traffic are tallied into packets and bytes which is scalable, a database of netflow information. Here the source address helps to identify the origin of netflow traffic and destination address, to find out the receiver. This information will be periodically sent to flow collector.

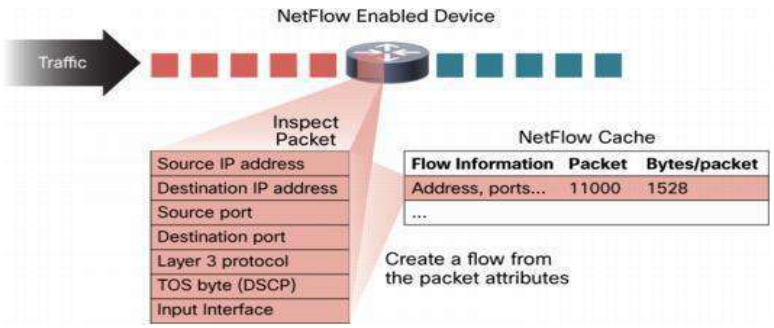


Fig. 3: NetFlow Cache [9].

Packets are classify into three categories where main concern for VOIP, Medium priority for VPN and low priority for others. The sampling is assigned based on their priority and collects through netflow cache.

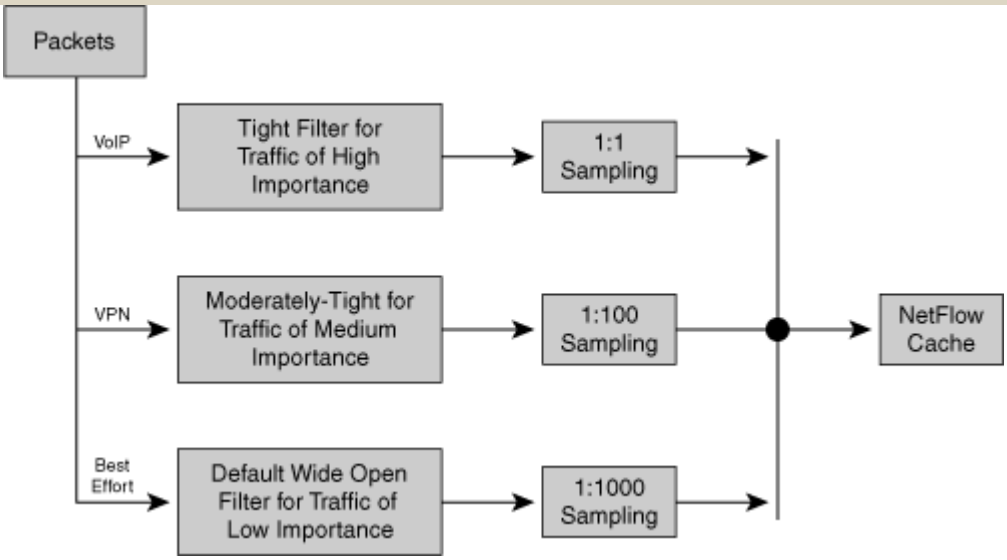


Fig. 4: Netflow Input Filters. [10]

Netflow collector

The Netflow collector [11] can be a storage source for network management. Here the routes and tracks of the packets are streamlined based on their enabled ports. These stored facts are used by enterprises for business applications like MPLS[12] and VPN traffic analysis, for billing based on their usage, analyzing the traffic in the flow, monitoring the reduction and deduction of unauthorized WAN traffic, analyzing new applications and their impact.

Netflow deployment

Netflow uses maximum 20% of CPU memory in the netflow cache.Netflow represents the traffic of flow upto 3% from the exporter to collector or end source. Actually the netflow analyzing will be like flow of vehicles on the highway. Netflow should be implemented with observation and maintenance towards every reciever.

Packet tracer simulator

Netflow is an indispensable tool [13] proposed by CISCO to gather information about the flow of network from the headquarters to the end receiver. Netflow version 9 is standardized in 2008 by IETF organization. The gathered information on netflow is viewed by the collector[15] through User Datagram Protocol (UDP) in netflow cache.

Packet Tracer is a proposed visual tool by Dennis frezzo [16] and his team at CISCO system. It acts as a platform in modern computer networks. It creates a drag and drop user interface in network topology. Packet tracer contains layer2 protocols such as Ethernet and PPP for tracing the routes of netflow. It also works with layer3 and layer4 protocols like IP, ICMP, ARP, TCP and UDP.

CISCO Packet Tracer[17] acts as an integral part of Networking Academy that stimulates programmes visually to the learners. It facilitates authoring, assessment and collaboration capabilities for the teachers and understanding some complex technical concepts.

Flows and packet lengths for all NetFlow export versions

Table1: NetFlow version

NetFlow Export Version Format	Number of Flows in a Packet	Packet Length (Bytes)
Version 1	24	1200
Version 5	30	1500
Version 7	27	1500
Version 8	51	1456

CPU utilization for a number of active flows

Sampled Netflow will significantly decrease CPU utilization to the router. On average sampled NetFlow 1:1000 packets will reduce CPU by 82% and 1:100 sampling packets reduce CPU by 75% on software platforms. The conclusion is sampled NetFlow is a significant factor in reducing Active flows.

IIOAB JOURNAL
ISSN: 0976-3104

Table2: Active Flows

Number of Active Flow in Cache	Additional CPU Utilization
10000	< 4 %
45000	< 12 %
65000	< 16 %

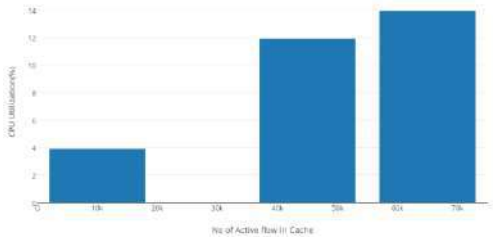


Fig. 5: CPU Utilization.

RESULTS

Statistics of a NetFlow cache

NetFlow Cache [18] is a forceful tool for analyzing the congestion in the netflow records. It helps to characterize the active flows from the source to the receiver. It facilitates necessity of flow as where to maximize and minimize i.e, the need to posses the small packets or large packets in the flow. This packet size and length of the flow in turn helps to report the congestion place and enable to solve the related issue easily and accurately.Below is the table about IP Packet distribution which displays the show ip cache flow in command in EXEC mode. The commands enable to code the statistics of ip cache flow in the network.

Table3: IP Packet Distribution

Range of Packets (Bytes)	Utilization (Percentage)
0-32	4.75%
33 - 64	10 %
65- 96	4.2 %
97-128	1.7 %
129-160	1.5%
161-192	0.9%
193-224	0.9%
225-256	0.9%
257-288	1.3%
289-320	3.0%

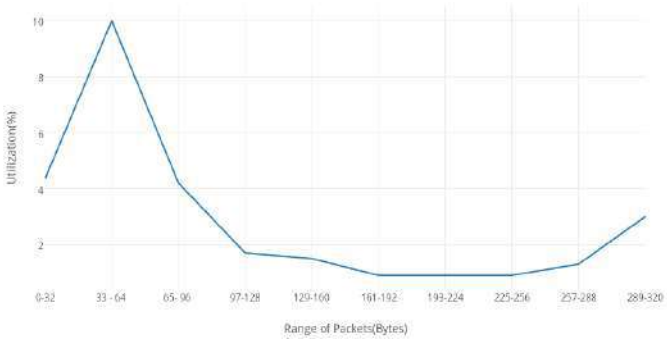


Fig. 6: Packet distribution.

ISSN: 0976-3104

Protocol statistics

NetFlow entries record the SYN Flag which is set in the first packet of each TCP connection. Using this information, it is possible to estimate accurately the number of active TCP flows in various aggregates.

Table 4: Protocol Statistics: TCP

Protocol	Total Flows	Flows /Sec	Packets/Flows	Packets/Sec	Active(Sec)/ Flow	Idle(Sec)/Flow
TCP-Telnet	2656854	4.3	86	372.3	49.6	27.6
TCP-FTP	5900082	9.5	9	86.8	11.4	33.1
TCP-WWW	5467782	887.3	12	11170.8	8.0	32.3
TCP-SMTP	25536863	41.4	21	876.5	10.0	31.3
TCP-BGP	24520	2.0	28	1.1	26.2	39.0

The above [Table 4] generates using the show ip cache flow command which enrich the administrator to investigate the large traffic. From the table TCP-BGP protocol used the greatest amount of network time is 39.0 sec when compare to other protocols. TCP- Telnet protocol uses the least amount of network time 27.6 sec. The Active flow is appreciable for TCP-Telnet with 49.6sec. But due to some congestion the active flow is not appreciable for TCP-www.

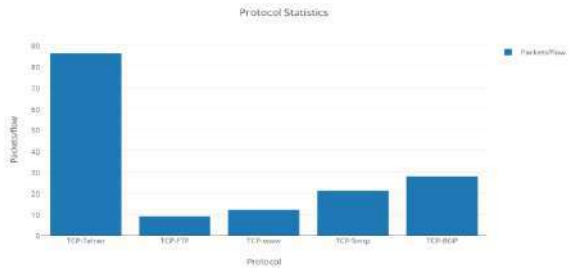


Fig. 7: Protocol Statistics – TCP.

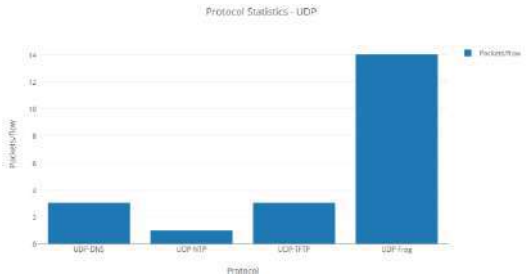


Fig. 8: Protocol Statistics – UDP.

Table 5: Protocol Statistics – UDP

Protocol	Total Flows	Flows /Sec	Packets/Flows	Packets/Sec	Active(Sec)/ Flow	Idle(Sec)/Flow
UDP-DNS	117240379	190.2	3	570.8	7.5	34.7
UDP-NTP	9378269	15.2	1	16.2	2.2	38.7
UDP-TFTP	8077	1.0	3	0.0	9.7	33.2
UDP-Frag	51161	1.0	14	1.2	11.0	39.4

The above [Table 5] shows UDP-DNS is balanced in all parameters especially in average number of flows and average number of packets per second. UDP-Frag utilizes more amount of network time as 39.34 sec when compare to other protocols. The average number of flows is not appreciable for UDP-TFTP and Frag. The average number of packets for the flows is only one packet per flow for UDP-NTP.

IIOAB JOURNAL
ISSN: 0976-3104

Top-Talkers

Top Talkers feature uses NetFlow functionality to obtain information regarding heaviest traffic patterns and most-used applications in the network. The Top Talkers can be sorted by either total number of packets or total number of bytes. Top Talkers help us to identify the heavily used parts of the system and assist in traffic study.

Table 6: Application based traffic

Application	Port Number	Bytes
TCP	06	727K
UDP	11	1095K

The above [Table 6] shows UDP traffic patterns are used mostly in the network when compare to TCP traffic patterns. This data helps us to plan our network. Additional information like interface, source and destination address are also retrieved.

NBAR protocol discovery

Network-Bases Application Recognition is an intelligent classification engine in Cisco IOS software that can recognize a wide variety of application, including Web-based and client/server applications.
NBAR includes a feature called Protocol Discovery. Protocol discovery provides an easy way to discover the application protocol packets that are passing through an interface.

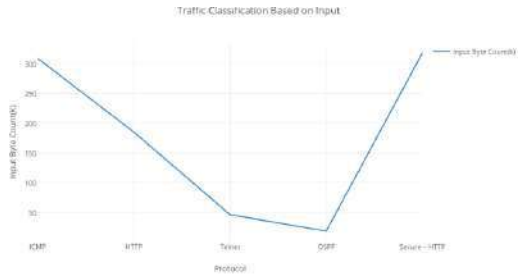


Fig. 9: Traffic classification – Input .

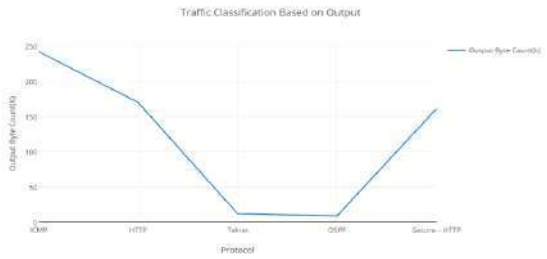


Fig. 10: Traffic classification – Output.

The above [Fig. 9] and [Fig. 10] helps us to determine the currently running protocol and applications on your network. Due to the congestion[19] in the network the amount of bytes that enter ingress and egress interface are not same. This [Fig. 9] and [Fig. 10] helps us to know the statistics for all interfaces on which protocol discovery is enabled. The Flow of HTTP protocol takes balanced byte count in input and output interface.

CONCLUSION

Network Management System using netflow data based monitoring feeds solution to many issues related to network congestion. In spite of this netflow cache plays unique role to insight over individual distribution of packets. However the need for capturing improves vision provided by netflow technology is certainly increasing day by day. In this paper, Analysis on a Personalized Network System through Netflow is presented which integrates CPU utilization, IP Packet Distribution, Protocol Statistics, Top Talkers and Protocol Discovery. Using this different command the network administrator can be able to analysis and control the congestion in the network thereby the utilization of networks like bandwidth, throughput, packet loss and active flows are managed in right manner. In future work, there will be a scope of performing real-time traffic analysis in high-speed networks.

ISSN: 0976-3104

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Rick Hofstede, Pavel C˘ eleda.[2014] Brian Trammell, Idilio Drago, Ramin Sadre, Anna Sperotto and Aiko Pras, *Flow Monitoring Explained: From Packet Capture to Data Analysis with Netflow and IPFIX*, IEEE Communications Surveys & Tutorials, 16(4).

[2] S. J. Murdoch and P. Zieli´nski, "Sampled traffic analysis by internetexchange-level adversaries," in In Privacy Enhancing Technologies (PET), LNCS. Springer, 2007.

[3] Cisco Sytems "NetFlow Services and Applications", White Paper,July 2002.

[4] NetflowArchitecture, http://www.viavisolutions.com/sites/default/files/technical-library- files/flowmonitoring-wp-nsd-tm-ae.pdf

[5] Cisco, Cisco IOS NetFlow Technology Data Sheet. <http://www.cisco.comlao/NetFlow>.

[6] S. Waldbusser, Remote network monitoring management information base. RFC2819/STD0059, *http://www.rfc-editor.org/*, May 2000.

[7] Manish khule, Megha Singh.[2015] *Tracking Low Grade Attack Using Cisco Packet Tracer Netflow,* International Journal of Emerging Technology and Advanced Engineering, 5(1)

[8] http://www.9tut.com/netflow-tutorial

[9] http://www.cisco.com/en/US/prod/collaletral/switches/ps5718/ps708/prod_white_paper0900aecd80673385.html

[10] http://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/netflow/nfwhite.html

[11] Rick Hofstede, Pavel C˘ eleda, Brian Trammell, Idilio Drago, Ramin Sadre, Anna Sperotto and Aiko Pras, *Flow Monitoring Explained: From Packet Capture to Data Analysis with Netflow and IPFIX*, IEEE Communications Surveys & Tutorials, Volume 16, Issue 4, 2014

[12] Gurwinder Singh, Er. Manuraj Moudgil, Comparative Analysis of MPLS Layer 2 VPN techniques,

[13] International journal of Computer Science Trends and Technology –3(4) Jul-Aug 2015.

[14] Cisco NetFlow, http://www.cisco.com/web/go/netflow

[15] B. Claise, "Cisco Systems NetFlow Services Export Version 9,"RFC 3954 Informational), Jul. 2008. [Online]. Available: http://www.ietf.org/rfc/rfc3954.txt

[16] Michael W Lucas. [2010] Network Flow Analysis. No Starch Press, Inc.

[17] Dragos Petcu, Bogdan Iancu, Adrian Peculea, Vasile Dadarlat and Emil Cebuc.[2013] Integrating Cisco Packet Tracer with Moodle platform Support for teaching and automatic evaluation, IEEE Xplore Networking in Education and Research.

[18] Garima Jain, Nasreen noorani, Nisha kiran, Sourabh Sharma.[2015] Designing & Simulation of Topology Network using Packet Tracer, International Research Journal of Engineering and Technology (IRJET), 02 (02).

[19] Shivam Choudary, Bhargav Srinivasan , Usage of Netflow in Security and Monitoring of Computer Networks , International Journal of Computer Applications (0975 – 8887), Volume 68– No.24, April 2013

[20] Gesu Thakur.[2013] Basics of causes and detection of congestion over TCP/IP networks, International Journal of Latest Research in Science and Technology, 2(1):609-611

[21] Yiming Gong, Detecting Worms, and Anomaly Activities withNetFlow.http://www.securityfocus.com/infocus/1796 >.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A STOCHASTIC SOFTWARE DEVELOPMENT PROCESS IMPROVEMENT MODEL TO IDENTIFY AND RESOLVE THE ANOMALIES IN SYSTEM DESIGN

Pratik Rajan Bhore*, Shashank D. Joshi, Naveenkumar Jayakumar

Dept of Computer Engineering, Bharati Vidyapeeth Deemed University College of Engineering, Pune, INDIA

ABSTRACT

Background: *It is observed that the existing research models vary within small scale industries considering the overall system design. Their tool focuses mostly on quality oriented questions and the manual or automated detection using only the source code for finding out the scope of improvement, but it misses out finding the real anomalies in the design phase.* ***Methods:*** *So it is a necessity to develop a model that can use the manual as well as automated detection also parameters such as anomaly density, reverse engineering and also use traceability matrix to trace from testing of a software code to the design requirements for identifying anomalies in the system design and its evolution. In this research, the authors present a semi-automated stochastic software development process improvement model which will identify and resolve the anomalies in system design and its evolution. The proposed model can be realized through the tool for the enhancement of software development process assessment in an economical manner consuming minimum resources. The proposed device encloses a set of questionnaires for phases of SDLC like Testing, Coding, Design and Requirement phases. These questions get designed because of the quality assessment and improvement. The tester answers the questions (KPA), and the result indicates the improvement needed in the considered KPA.* ***Results:*** *Moreover, hence the authors use the model to detect and resolve such software design anomalies and also calculate the anomaly density along with the bug %.* ***Conclusion:*** *The proposed model tackles these design anomalies and enhance the quality of the software by manually as well as automatically testing and finding the more anomalies in the software. Also, it presents a better software development process model integrated with TDD agile model which is novel than the other existing models.*

KEYWORDS
, Design Anomalies; Software Quality; Software Testing; Process Improvement; Software Development; Software Engineering

Received: 5 June 2017
Revised: 14 June 2017
Published: 4 July 2017

***Corresponding Author**
Email:
pratik4u2007@gmail.com

INTRODUCTION

In the previous articles [1] by the authors of this research, the first article consisted of the novel approach that what were the basic parameters [1] that were being considered by the authors to create the model. The basic parameters got refined, and more parameters were added in as the research progressed. Then the next article explained about the unique methodology and solution which was going to be used for tackling with the design anomalies in software [2]. In this paper, the authors present the entire implementation and result in the analysis where the explanation is done about how the semi-automated stochastic model works. These results will fulfill the main objectives of quality enhancement as well as the identification and resolving of anomalies in the system design. Below take a look once again at the proposed stochastic model and the workflow of tasks carried by developer or tester in the architecture.

So before going into the methods for implementation of the model here, the proposed stochastic model architecture is shown here. From the previous article [2] about the proposed architecture, the author integrates that model into the agile model Test Driven Development [3]. In the TDD the focus is mainly on testing and refactoring of coding. The author takes the TDD further into the design phase from the testing and coding phases. It is how the newly integrated semi-automated stochastic model looks when integrated with TDD to give a better software development process improvement model [3]. The author hence proposes a new model for the developers or testers that can be used to test the software and develop it. It consists of four phases namely: Testing, Coding, Design, and Requirement. The earlier models all work only till the refactoring of code while the stochastic model can use reverse engineering to trace back to the design anomalies and also use traceability matrix to trace back from test cases in testing to requirements. It is the novelty of this model which is more than what the earlier models have to offer. Below the stochastic model architecture is seen in [Fig. 1].

Here above it is observed how the entire proposed model works that what is its actual workflow of the model. It is integrated with the TDD agile model to give a better process improvement model to the developer and tester for the software development and testing processes.

To give a proper introspect into the working of this model and to support the research the authors shows exactly how the flow happens.

Firstly the developer/ tester select the software he wants to test. Here the authors have tested three software namely TIC TAC TOE, Pokémon Attack, and ScratchPad. All these are mobile applications which were tested. The testing tool is used for checking if any bugs [4] can be resolved from the outside without going within the system design. Then if the bugs are not fixed from the outside, then coding phase is entered for testing. The code is tested for anomalies firstly. The anomalies are identified here from the code. The functional testing shows that there are code and design anomalies within the software [4] so explicitly testing these phases helps in eliminating these anomalies. The code [5] consists of the anomaly

IIOAB JOURNAL
ISSN: 0976-3104

which can be checked using the UML tool [5] that if this anomaly affects the design directly. The anomaly in the code [6] directly makes design anomalies occur. It is a part of the entire design anomaly that occurs. There are various types of these anomalies such as Blob, Functional Decomposition, and Spaghetti Code [6]. The proposed model helps in mapping the anomaly from testing code to the design of the software. It helps in eliminating the actual anomaly from the design phase. The occurring anomaly is code can be mapped in the design diagram of the software. The model supports the elimination from directly the design phase using reverse engineering [7] if the resolving of the anomalies from code [7] does not eliminate the entire anomaly in design. The design consists of the requirements as well in the use case that are gathered before creating the design. The traceability matrix document helps in tracing back and forward from the test cases in testing to the requirements in the design [8]. It helps in mapping the design anomaly as well and determining the relation between the phases. The testing and the mapping using the traceability matrix are all manuals which assist in identifying the testing or the bug % [8] in the software. This manual process is then overtaken by the automated process of finding the anomaly density [9] testing the code and design. In the end, the quality is enhanced by eliminating the anomalies. Having a semi-automated model is better than only having a manual or only an automated model. It is said because manual detection takes alot of time for detection of the anomalies while automated models consume alot of time in calculating false positives. Hence a semi-automated approach is better to have in developing and testing the software.

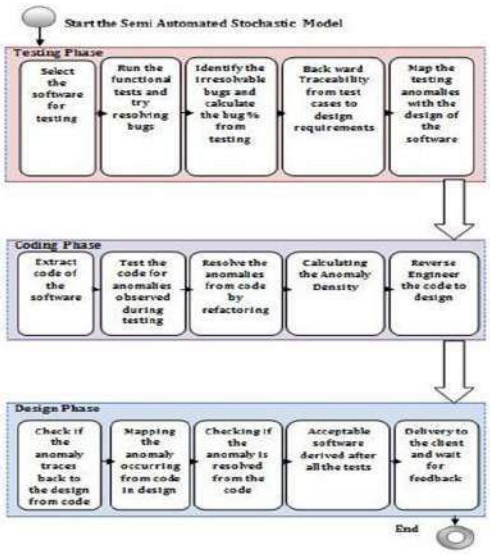


Fig. 1: Stochastic model derived after integration with TDD Agile model

MATERIALS AND METHODS

The authors implement this module on the three different software namely: TIC TAC TOE, Pokémon Attack, and ScratchPad. These three are mobile applications in use. Testing is done on this software firstly to correct the bugs that can be corrected from the outside. If the bugs or errors cannot be corrected from the outside, then the problem lies in the code and design. Testing helps in mapping the anomaly [9] and relating it to the design phase. This module is a semi-automated where the testing part of the code and design of software is all automated while the mapping of the anomalies is all done manually after it is identified by the module.

So firstly by using the testing tool, the software are tested for any bugs or errors that cannot be solved. The test cases are created based on these tests that are carried by the tester or developer. The bug % or the testing % is determined by calculating it. The parameters used to calculate it is the bug or error scenarios of the testing carried and the test cases.

Let's have a proper look at calculating the bug%. Consider n as the total no. of test cases of the software, y as the bug or error scenarios and x as the other scenarios which consist no bugs or resolvable bugs. Hence the equation derived from this is [10]:

$$\text{Bug\% /Error\%} = \frac{\text{the total no. of bug scenarios (y)}}{\text{the total no. of test cases (n)}} \times 100$$

Functional testing the TIC TAC TOE software and finding Bug Percentage

Below here the testing tool shows the no. of test cases done on the TIC TAC TOE software. It consists of 11 test cases (n), and the no. of unresolved bugs [10] using the testing tool are 3 (y) here. It indicates the

IIOAB JOURNAL
ISSN: 0976-3104

problem lies within the code and design of the software. The test casesof the software hence can be seen below [Fig. 2].

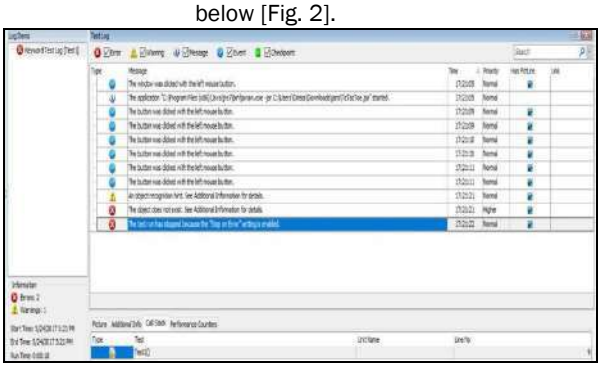


Fig. 2: Test cases of TIC TAC TOE software

By clicking on the bugs [10], it is given in the design of the software that there is an anomaly affecting one of the buttons of the software. It can be mapped in the design phase as well. Hence, it proves the occurrence of design anomaly in the software. The bugs identified hence can be seen below [Fig. 3].

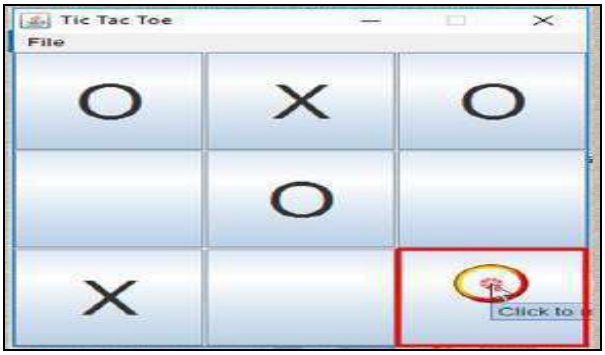


Fig. 3: Bug detected during functional testing in TIC TAC TOE

Using the equation mentioned above i.e. y/n x 100 gives the bug or the error % [10] in the software. Hence using it, we calculate the % which gives:

Bug %: 3/11 x 100 = 27.2 %

It is the bug % [10] in the software application which is found.
Similarly, the other two software are tested as well which are Pokémon Attack and ScratchPad.

Testing the Pokémon Attack software

Here similarly to the TIC TAC TOE, the testing tool shows the no. of test cases done on the Pokémon Attack software. It consists of 17 test cases (n), and the no. of unresolved bugs using the testing tool are 2 (y) here. It indicates the problem lies within the code and design of the software. By clicking on the bugs, it is given in the design of the software that there is an anomaly affecting the GUI [11] touch and the run button as well of the software. It can be mapped in the design phase as well. Hence, it proves the occurrence of design anomaly in the software.

Using the equation mentioned above i.e. y/n x 100 gives the bug or the error % in the software. Hence using it, we calculate the % which gives:

Bug %: 2/17 x 100 = 11.7 %

It is the bug % in the software application which is found.

Testing the scratchpad software

Similarly here the testing tool shows the no. of test cases done on the ScratchPad software. It consists of 12 test cases (n), and the no. of unresolved bugs using the testing tool are 4 (y) here. It indicates the problem lies within the code and design of the software. By clicking on the bugs, it is given in the design of the software that there is an anomaly affecting the open file option on the GUI [11] design of the software. It can be mapped in the design phase as well. Hence, it proves the occurrence of design anomaly in the software.

IIOAB JOURNAL
ISSN: 0976-3104

Using the equation mentioned above i.e. y/n x 100 gives the bug or the error % in the software. Hence using it, we calculate the % which gives:

Bug %: 4/12 x 100 = 33.3 %

It is the bug % in the software application which is found.

Testing the code and design of the software and finding Anomaly Density

After determining that there are anomalies in the software within then, the tester carries tests on the code and design of the three software. Based on the parameters in the code, the anomaly density is calculated here. Anomaly Density [12] is an important parameter for calculating the quality of the software. The anomaly density is the total no. of anomalies found in all the modules of the software upon the total no. of the line of code. Then, later on, it can be converted to KLOC which thousand line of code [13].

$$\text{Anomaly Density\%} = \frac{\text{Total no. of anomalies}}{\text{Total Size (KLOC)}}$$

Firstly in the TIC TAC TOE software, the code is tested where the error seen in the functional test case was viewed. Below it is shown how the anomaly lies in the design of the software and it is the button of the application. The identified anomalies in the code of the software hence can be seen below [Fig. 4].

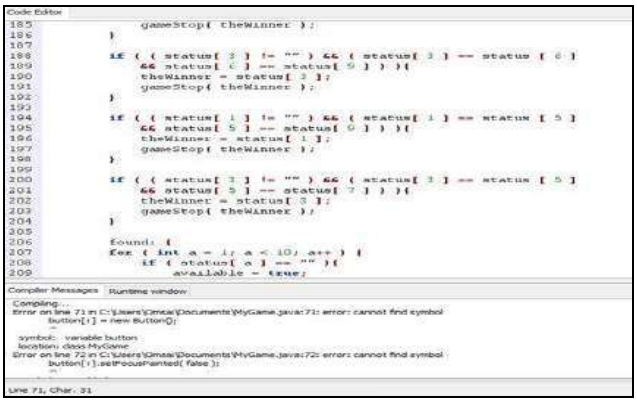


Fig. 4: Anomalies identified in the code of TIC TAC TOE

Further, the code is used to reverse engineer [14] to the design phase of the software where it is seen the design anomaly called as "Blob" has been created. A blob consists of a God class that stores all the attributes and operations and the subclasses are not seen as they are not functional. The identified anomaly blob hence can be seen below [Fig. 5].

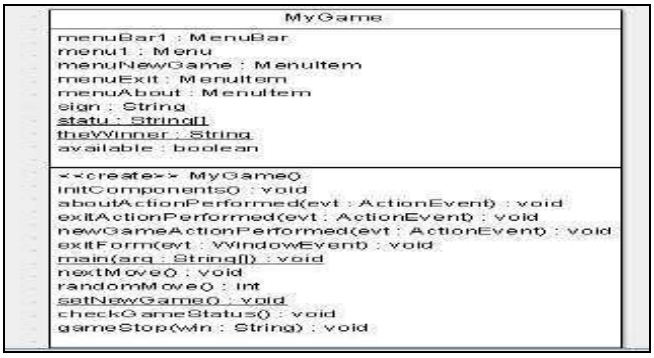


Fig. 5: Blob anomaly Identified in the OO Design of TIC TAC TOE

After the anomaly is resolved from the code, it is visible there is no anomaly in the code now. The proposed module helps in deriving this conclusion that there is no anomaly in the code now as seen below [Fig. 6].

IIOAB JOURNAL
ISSN: 0976-3104

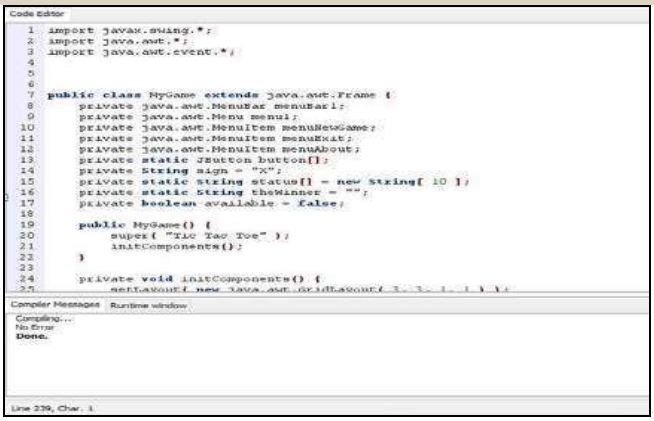


Fig. 6: Code refactored using the module

After eliminating the anomalies, the developer uses the reverse engineering tool [15] to check the OOD [15] of the software in design phase if the design anomaly is eliminated and it gets disposed of in the design like seen below [Fig. 7].

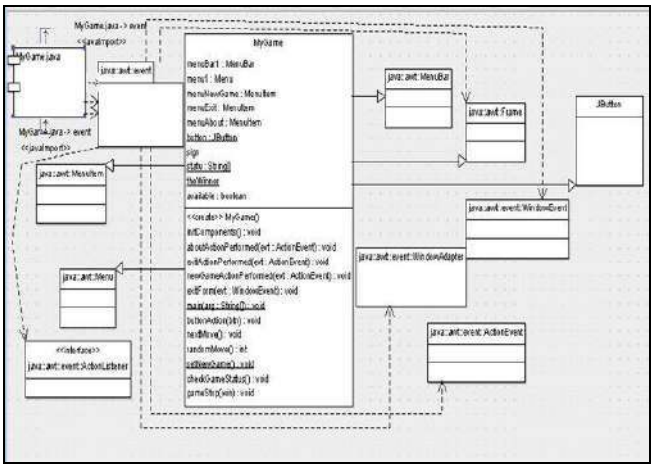


Fig. 7: Design anomaly resolved (Reverse Engineering)

Similarly the other two software namely Pokémon Attack and ScratchPad as well are tested. The anomalies are found in the code and design as well of both the software. In the Pokémon software is the anomalies are found mapped from code to the design also seen during functional testing of the software. The mapping is done manually using these results in testing, coding, and design. The anomaly in the interface [16] touch of the gaming application, as well as the run button, is seen here in the code and design as well. The anomaly occurring in the design is the functional decomposition. None of the functions are seen as there is no relationship [16] determined in the OOD [16] of the software. It shows the design anomaly in the software that is why while playing the game there is a problem in the design interface as there is a functional anomaly in design. It is directly mapped from the test case to the design requirements using the traceability matrix parameter. By using the proposed module, this anomaly is resolved just like in the TIC TAC TOE example shown above.☐

Further, the testing of the final software of ScratchPad is done as well. The code and design are tested here. The anomaly called Spaghetti Code is observed in this software. This anomaly seen during functional testing made the open file option disappear. The code gets tangled due to this and directly affects the main class in the OOD [17] of the software. The code gets tampered with, and major parts are missing. Hence it gets identified in the proposed module. The same anomaly is mapped seen in functional testing to the design. The open file anomaly occurred due to spaghetti code affects the OOD as well. In the main class, this operation and attribute are missing hence affecting the interface of the software as well. The design is the phase the customer comes in contact with hence it is important to keep it anomaly free.

Just like the first software is resolved from anomalies, in the same manner, the other two software are also made anomaly free as all the design anomalies get resolved. The tests result in identifying three major types of anomalies in the software namely Blob, Functional Decomposition, and Spaghetti Code. Also, the manual mapping is done of this in the design of the software from the test cases and code.

Hence proving the semi-automated approach carried out by authors. Further, the anomaly density [18] is calculated by the three software.

IIOAB JOURNAL
ISSN: 0976-3104

Calculating the Anomaly Density for TIC TAC TOE software here,
Anomaly Identified= 11
Line of Code=237

Anomaly Density %: 11/237 LOC x 1000(KLOC) = 46.4%

Calculating the Anomaly Density for Pokémon Attack software here,
Anomaly Identified= 16
Line of Code= 231

Anomaly Density %: 16/231 LOC x 1000(KLOC) = 69.2%

Calculating the Anomaly Density for ScratchPad software here,
Anomaly Identified= 5
Line of Code= 113

Anomaly Density %: 5/113 LOC x 1000(KLOC) = 44.2%

The Anomaly Density % [19] in all the three software is more than the calculated Bug % [20]. It proves that the anomalies need to be addressed by testing the code and design phases of the software then only the functional black box testing of it. The normal outside testing detects few bugs and indicates there are many more anomalies within which cannot be resolved using the testing tool. Hence using the proposed module the code and design are tested, and the anomalies are identified as well as resolved. Hence the Anomaly Density % is more than the Bug %. Anomaly Density is a parameter used to calculate the Quality of the Software. If the anomaly density is more than bug %, then there are more anomalies identified by using the stochastic model. Hence the proposed module is better for testing the software and resolving the anomalies within than the existing models.

The author's module also provides the additional features of HTML editor rather known as IDE Expert which can be used to correct the dynamic code and designs. The tests are carried on this extra feature where the module detects the anomaly from the code. The suggestion is shown where the anomaly lies in the code which can resolve anomalies in the design.

RESULTS

In the result analysis, the expected results have been derived after conducting the necessary tests to support the proposed research. The bug percentage is calculated firstly while functional testing and later the anomaly density is calculated when the tester tests the code and design of the software. Firstly the Software Stats are shown where the software on which the tests are carried is seen, and then the no of classes which are part of the code and no of the line of code is shown. These classes from code are directly related to the classes in Object Oriented diagrams. It is how anomalies like Blob, Functional Decomposition, and Spaghetti Code occur in the design. This model resolves such anomalies in the design. The software statistics hence can be seen below [Table 1].

Table 1: Software statistics

SOFTWARE APPLICATION	CLASSES	LINE OF CODE
Tic Tac Toe	108	237
Pokémon Attack	433	231
ScratchPad	79	113

Later the bug percentage [21] is derived from functional black box testing which is done manually to the each software to solve the bugs that can be solved from outside and also determine the bugs that need to be solved using white box testing from the within of code and design. Here the software functions are tested where unit components and the integration of these components is all tested. Test cases are written here by the developer or tester as the testing is carried. Later the bug percentage is calculated based on the test cases carried and the bug scenario detected that are irresolvable from the outside. The bug percentage results hence can be seen below [Table 2].□

Table 2: Bug percentage

SOFTWARE APPLICATION	TEST CASES	BUG SCENARIO
Tic Tac Toe	11	3
Pokémon Attack	17	2
ScratchPad	12	4

IIOAB JOURNAL
ISSN: 0976-3104

Next is the calculation of the anomaly density which is the parameter that shows the quality of the software. It indicates that the anomalies which are resolved to give quality enhancement [21]. The LOC as well is the parameter used here. It is carried in an automated way using the model. It is calculated based on code and design of the software. The anomaly density results hence can be seen below [Table 3].

Table 3: Anomaly density

SOFTWARE APPLICATION	ANOMALY IDENTIFIED	LINE OF CODE
Tic Tac Toe	11	237
Pokémon Attack	16	231
ScratchPad	5	113

The final results display the bug % and the anomaly density % which is more than it. It shows that during the testing of code and design more anomalies will be identified and resolved more than by only functional testing. Hence the manual and automated way resolves more design anomalies than by only functional testing. The model combines to give out the semi-automated model which will eliminate all the anomalies manually and automatically to give better results as well as give better quality. It will be beneficial to give the customer a better experience and to get better feedback as well. Also, the design anomalies will be resolved, and the quality will improve as well. The final results hence can be seen below [Table 4].

Table 4: Final results

SOFTWARE APPLICATION	BUG % (TESTING)	ANOMALY DENSITY %
Tic Tac Toe	27.2%	46.4%
Pokémon Attack	11.7%	69.2%
ScratchPad	33.3%	44.2%

Comparing manual and automated testing carried in the semi-automated stochastic model for identifying and resolving design anomalies. The graph hence can be seen below [Fig. 8].

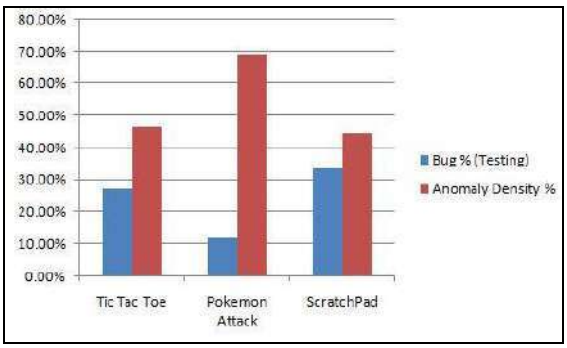


Fig. 8: Graph comparison between manual and automated testing

CONCLUSION AND FUTURE WORK

In the entire research, the authors have proposed a stochastic software development process improvement model the main aim of this system was to enhance quality in various software by identifying the anomalies from the code and design phases and resolving them. In this study, the authors fulfill the objectives by detecting and correcting these anomalies from the code and employ the technique of reverse engineering and traceability matrix [22] to check whether the actual design anomaly has been removed entirely. The customer using the software is in contact directly with the design phase when he uses it. It is the interface between the full software and the customer. That is why this model tries to improve the quality of the system design for the satisfaction of the client and to give a better feedback in return to the developer. These anomalies try to compromise the performance of the software. Also in today's world as there is much competition between software organizations everyone would like their software to be used more than other and to be bug-free. The proposed model also has an additional feature where it helps in achieving this goal where most of the online systems employ the dynamic language of HTML while most of the offline usable software employ the JAVA language source code in implementing their software and also use object oriented programming to form their design. The model deals with such do not and give them a chance to improve their quality with less time wastage and also save a lot of companies' resources. Also, it helps in creating highly reliable software.

IIOAB JOURNAL
ISSN: 0976-3104

Shortly, many researchers can take this work forward by addressing the initial stage of the requirement stage as well. In this research work, the authors address the design and use case requirements as well as its relationship to the coding phase. In the coming days, many researchers can try to establish the relationship of the actual requirement phase with the design phase as it is the first phase of the software development cycle. Here the employed agile models such as TDD and FDD that focus on the changes that can be brought to the code to remove design anomalies but in the future researchers can integrate their model with agile models like XP or Scrum that focus on the requirements more than the code where there are breakpoints to adapt to changes requirements. It can further improve the quality of the software as well as the proposed model further as the technology is changing with the coming days [23]. This further work in this area will be beneficial further to many more organizations as well.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Bhore PR, Joshi SD, Jayakumar N. [2016] A Survey on the Anomalies in System Design: A Novel Approach. International Journal of Control Theory and Applications, 9(44):443-455.
[2] Bhore PR, Joshi SD, Jayakumar N. [2017] Handling Anomalies in the System Design: A Unique Methodology and Solution. International Journal of Computer Science Trends and Technology, 5(2):409-413.
[3] Jayakumar N, Bhor M, Joshi SD. [2011] A Self Process Improvement For Achieving High Software Quality. International Journal of Computer Science Trends and Technology, 3(5):306-310.
[4] Patil TB, Joshi SD. [2015] Software Improvement Model for small scale IT Industry. International Journal of Advanced Research in Computer and Communication Engineering, 4(5):601-604.
[5] Fenton N, Pfleeger SL. [1998] Software Metrics: A Rigorous and Practical Approach. ACM Transactions, 15(2):126-139
[6] Fowler M. [1999] Refactoring—improving the design of existing code. 1st edn. Addison-Wesley, Reading IEEE.
[7] Brown WJ, Malveau RC, Brown WH, McCormick HW, Mowbray TJ. [1998] Anti Patterns: Refactoring Software, Architectures, and Projects in Crisis. 1st edn. Wiley, New York IEEE, 15(2):101-112
[8] Kessentini M, Kessentini W, Sahraoui H, Boukadoum M, Ouni A. [2011] Design anomalies Detection and Correction by Example. IEEE.
[9] Yoshida N, Saika T, Choi E, Ouni A, Inoue K. [2016] Revisiting the relationship between code smells and refactoring. IEEE 24th International Conference on Program Comprehension (ICPC), 33(1):1-4.
[10] Mens T, Tourwé T. [2004] A Survey of Software Refactoring. IEEE Transactions on Software Engineering, 30(2):126-139.
[11] Sahraoui H, Abdeen H, Bali K, Dufour B. [2015] Learning dependency-based change impact predictors using independent change histories. Information & Software Technology IEEE, 23(2):220-235.
[12] Moha N, Gueheneuc YG, Duchien L. [2010] DECOR: A Method for the Specification and Detection of Code and Design Smells. IEEE Trans. Softw. Eng.
[13] Martin R. [1994] OO Design Quality Metrics An analysis of dependencies. IEEE.
[14] Chechik M, Gannon J. [2001] Automatic Analysis of Consistency between Requirements and Designs. IEEE Trans. Softw. Eng.
[15] Fenton N, Ohlsson N. [2000] Quanti. Analysis of Fault & Failure in Complex Softw. Sys. IEEE Trans. Softw. Eng.
[16] McCabe TJ. [1976] A Complexity Metric. IEEE Transactions on Software Engineering, 2(4):308-320.
[17] Marinescu R. [2004] Detection Strategies Metrics based rules for detecting design flaws. IEEE Computer Society.
[18] Dhambri K, Sahraoui H, Poulin P. [2008] Visual detection of design anomalies. IEEE Computer Society.
[19] Abdeen H, Shata O. [2012] Metrics for Assessing the Design of Software Interfaces. Int. J. of Adv. Res. Comput. Commun. Eng., 1(10):1-8, IEEE.
[20] Hossain A, Kashem A, Sultana S. [2013] Enhancing Software Quality Using Agile Techniques. Springer.
[21] Gautam AK, Diwekar S. [2012] Automatic Detection of Software Design Patterns from Reverse Engineering. Springer.
[22] Bhore PR. [2016] A Survey on Storage Virtualization and its Levels along with the Benefits and Limitations. International Journal of Computer Sciences and Engineering, 4(7):115-121.
[23] Bhore PR. [2016] A Survey on Nanorobotics Technology. International Journal of Computer Science & Engineering Technology, 7(9):415-422.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

TOWARDS OPTIMAL ALLOCATION OF RESOURCES IN CLOUD MODIFIED MAPREDUCE USING GENETIC ALGORITHM

D. Arivudainambi[1], D. Dhanya[2*]

[1]*Department of Mathematics, Anna University, Chennai, INDIA*

[2]*Research Scholar, Anna University, Chennai, INDIA*

ABSTRACT

Background: *Cloud computing is one among the unfathomable knowledge exploration in the field of Information Technology. A virgin as well as a task emerging area of study, cloud computing provides an innovative model for the distribution of multiple applications in various organizations.* ***Methods:*** *The development and growing popularity of cloud computing indicates the evolution in the way IT infrastructure and services are distributed and expended. This increased use of cloud computing resulted in resource problems which have to be solved for better usage of the clouds. In this paper, we present an efficient method for solving resource problems in the cloud using modified map reducing algorithm.* ***Results:*** *Here we employed a Map Reduce programming model with GA for distributed parallel computing and execution on a virtualization process which is used to detect non-sufficient reductions in the execution time and to detect the decrease in the computing time.* ***Conclusions:*** *GA is compared with existing methods such as Clustering large applications(CLARA), Partioning around medoids(PAM), Clustering large applications based on randomized search(CLARANS) in which the simulation results of GA aim on improving the solutions both in execution and computation.*

KEY WORDS
Cloud computing, MapReduce, resource allocation, genetic algorithm, clustering algorithms.

Received: 18 March 2017
Accepted: 30 May 2017
Published: 5 July 2017

***Corresponding Author**
Email:
dhanya9870@gmail.com

INTRODUCTION

Cloud computing is a large scale distributed computing paradigm in which a pool of computing resources are available to the users via the Internet [1]. Cloud computing is a metaphor used by technology or IT Services companies for the delivery of computing requirements as a service to a homogeneous community of end-recipients. Recently, the cloud computing has been gaining lots of interests from researchers and IT industry due to its many abilities to provide flexible, dynamic IT infrastructures, QoS guaranteed computing environments, configurable software services and the state-of-the-art enabling cloud computing technologies [11]. Cloud computing clusters provide a large scale computing environment for scientific users. However, a large scale biological application often involves various types of computational tasks which can benefit from different types of computing clusters [9]. The use of cloud computing technology has gained popularity in recent years, and many companies are currently moving their business to the cloud, by deploying their services and executing their workloads in private or public clouds according to the application requirements or the particular business models [2].

In recent years, the application of high performance and distributed computing in scientific practice has become increasingly widespread. Among the most widely available platforms to scientists are clusters, grids, and cloud systems [10]. A storage cloud provides storage services, while a compute cloud provides computer services [13] It is important to note that we are implicitly considering the possibility of a hybrid deployment, i.e., the resources can be placed in different clouds, this multi-cloud setup can be suitable for deployment of independent virtual resources or for loosely-coupled multi-component services with no or weak communication requirements [3]. Data analytics play a key role in planning, problem solving, and decision support tasks. Data analytics applications typically process large amounts of data from both operational and historical data sources and the processing is primarily read-only with occasional batch inserts [5].

Scientific computing is a field of study that applies computer science to solve typical scientific problems. It is usually associated with large scale computer modeling and simulation and often requires a large amount of computer resource [7]. Providers such as Amazon, Google, Sales force, IBM, Microsoft, and Sun microsystems have begun to establish new data centers for hosting cloud computing applications in various locations around the world to provide redundancy and ensure reliability in case of site failures [4]. Many task computing (MTC) is novel paradigm, which tries to solve the problem of executing multiple parallel activities in multiple processors. Multiple processors may refer to a large cluster or a cloud [12]. The cloud storage service (CSS) relieves the burden of storage management and maintenance. However, if such an important service is vulnerable to attacks or failures, it would bring irretrievable losses to users since their data or archives are stored into an uncertain storage pool outside the enterprises [6].

Map Reduce is a programming model used for processing large data sets in a highly-parallel way. Users specify the computation in terms of a "map" function that processes a key/value pair to generate a set of intermediate key/value pairs, and a "reduce" function that merges all intermediate values associated with the same intermediate key [18]. It is one of the most popular programming models designed to support the development of such applications [16]. Map-Reduce is attracting a lot of attention, proving both a source for inspiration as well as the target of polemic by prominent researchers in databases. In database terms, map reduce is a simple yet powerful execution engine, which can be complemented with other data

storage and management components as necessary [14]. The map-reduce paradigm has been drawing the attention of the computational biology community particularly in the last year. However, majority of these applications are limited to simply distributing the data which are then handled by existing sequential software [20].

Hadoop is a popular open source cloud computing framework that has shown to perform well in various usage scenarios. Its Map reduce framework that offers trans-parent distribution of compute tasks and data with optimized data locality and task level fault tolerance; its distributed file system (DFS) offers a single global interface to access data from anywhere with data replication for fault tolerance [15]. On the other hand, data analysts in most companies, research institutes, and government agencies have no luxury to access large private Hadoop/ Map Reduce clouds. Therefore, running Hadoop/ Map Reduce on top of the public cloud has become a realistic option for most users [19]. One of the biggest merits of cloud storage is that users can access data in a cloud anytime and anywhere, using any device [8]. The simplicity of map reduce, its wide-spread usage, and its ability is in capturing the primary challenges of developing distributed applications [17].

In [21] Semantic web an emerging area to augment human reasoning has been proposed. Various technologies were being developed in those arenas which have been standardized by the World Wide Web Consortium (W3C). One such standard was the Resource Description Framework (RDF). Semantic web technologies can be utilized to build efficient and scalable systems for cloud computing. Hadoop's mapreduce framework is used to answer the queries. The results show that they can store large RDF graphs in hadoop clusters built with cheap commodity class hardware. Furthermore, their framework was scalable and efficient and can handle large amounts of RDF data, unlike traditional approaches.

In [8] storage in the cloud is proposed with the SPKS scheme for cloud storage services. It allows the CSP to participate in the decipherment, so that the user could pay less computational overhead for decryption. Furthermore, it was searchable encryption scheme; the CSP could search the encrypted files efficiently without leaking any information. It was probable that the proposed schemes have semantic security against adaptive chosen plaintext attacks.

In [12] the performance of Hadoop is been analyzed an implementation of mapreduce programming model for distributed parallel computing, executing on a virtualization environment comprised of 1 + 16 key pair values running the VMW. A set of experiments using the standard Hadoop benchmarks has been designed in order to determine whether or not significant reductions in the execution time of computations are experienced on the virtualization platform on a departmental cloud. The results mainly focused on the computation time. In [23] a cloud-computing based evolutionary algorithm is been presented using a synchronous storage service as a pool for exchanging information among the population of solutions. The multi-computer was composed of several normal PCs or laptops connected via Wi-Fi or Ethernet. The effect of how the distributed evolutionary algorithm reached the solution when PCs was tested whether that effect also translates to the algorithmic performance of the algorithm. To this end different problems were addressed using the proposed multi-computer, analyzing the effects that the automatic load-balancing and synchronization had on the speed of algorithm successful, and analyzing how the number of evaluations per second increases when the multi-computer includes new nodes.

The paper is organized as follows. Section 2 summarizes some of the formal definitions of the problem, section 3 presents the related work, section 4 presents the genetic algorithm proposed in this paper for solving the resource problem in the clouds using the map reduction algorithm, section 5 contains the experimental part of the paper where the performance of the proposed approach is evaluated and finally the conclusions are drawn in section 6.

RELATED WORK

Many definitions exist resource management techniques for cloud networks available in the literature.
K. Meri et al. [24] illustrated cloud-based evolutionary algorithms: An algorithmic study. This work shows the effect of how distributed evolutionary algorithm reaches the solution when new PC are added and tested and whether the effect translates the algorithmic performance. Different problems were addressed by analyzing the effect of load balancing and synchronization and are measured by number of evaluation per second when new nodes are added. The Evolutionary parallel algorithm solves the proposed problem and it was feasible both in homogeneous and heterogeneous environment.

Kai Zhu et al. [25] proposed Hybrid Genetic Algorithm for Cloud Computing Applications. An effective load balancing strategy improves the task throughput of cloud computing. Virtual machines are selected as a fundamental processing unit of cloud computing where the resources increases and vary energetically by the utilization of virtualization technology. As a result performance of load balancing has become complicated and is complex to be achieved. Multi-agent genetic algorithm (MAGA) a hybrid algorithm of GA was proposed whose performance is far superior to that of the traditional GA. The experimental results of paper shows the advantage of MAGA over traditional GA, where load balancing problem is solved in cloud computing moreover better performance is achieved.

Shin-ichi Kuribayashi et al. [26] proposed the Optimal Joint Multiple Resource Allocation Method for Cloud Computing Environments. This research develops a resource allocation model where both processing

ability and bandwidth are allocated simultaneously to each service request and rented out on an hourly basis. The allocated resources are dedicated to each service request. An optimal joint multiple resource allocation method is proposed based on the above resource allocation model where it reduces the request loss probability and as a result, reduces the total resource required, compared with the conventional allocation method. The simulation result shows that the proposed method allocates resources fairly among multiple users efficiently.

Fang Yu et al. [27] illustrated Quantitative Analysis of Cloud-based Streaming Services. As online streaming services significantly increase in recent years, it becomes a critical issue to offer quality service via systems that benefit from cloud computing developments. The study on quantitative analysis to estimate and further improve the quality of cloud-based streaming services, deriving theoretical results on operation characteristics of queuing models with mild assumptions. By simulating the continuous-time Markov chain, according to the adopted operations rules for VMs, we also get performance indicators such as the average lag time and interruptions that a customer may experience under different environmental settings.

PROPOSED METHODOLOGY FOR SOLVING RESOURCE PROBLEMS IN A CLOUD

Cloud computing is a rapidly developing area which offers a possible increase in the flexibility and efficiency of the service providers. With the help of cloud computing data's can be provided through shared computing resources and can be easily accessed through the internet. These include applications, development tools and servers. Amidst of all these advantages cloud offers, the major drawback has been the resource problem. The major intention is to develop a system which can solve the resource problems in the clouds. For this purpose we have used map reducing algorithm which proves to be an efficient method for reducing the over usage of resources such as execution and computing time.

In our proposed method, we implemented a Map Reduce Programming model for distributed parallel computing and execution on a virtualization process which is used to detect non-sufficient reductions in the execution time and to detect the decrease in the computing time. For an efficient mapper, Genetic Algorithm (GA) is used for optimizing the parameters of map reduce algorithm. A map and reduce algorithm is used to abstract the complex parallel computations. The Apache Hadoop model is used for map reduce. The proposed method is explained in detail in the below sections.

The HADOOP distributed file system stores a large amount of datasets and also reduces the Map Reduce jobs by computing the clusters within the key pair values. HDFS stores files across a collection of servers in a cluster. Files are decomposed into blocks, and each block is written to more than one of the servers. HDFS ensures data availability by continuously monitoring the servers in a cluster and the blocks that they manage. The key pair values in Hadoop classified as Name key pair values and Data key pair values. Name key pair values will allocate a data request to data key pair values from the master to the workers. The Name key pair value monitors jobs through the Job Tracker process, and the name key pair values execute and track tasks through Task tracker. Name key pair value can be replicated to avoid single point failure. HDFS deliver inexpensive, reliable, and available file storage. But the major component of Hadoop is the parallel data processing system called Map Reduce.

Hadoop Database for Proposed Method

The database forms the basic part in the map reducing process. In our proposed method, the database is the various documents that are selected for performing the map reduce programming. The databases are selected such that the word document contains the more frequent words at regular interval so that the map reduce framework can be effectively performed. We extract the keywords from the documents and based on their frequency, the sets are formed. Then, we compute the relevancy between these keyword sets. Keyword-based relevancy measure relies on the idea that the content of a document can be characterized by a set of keywords that is a set of words expressing the most significant concepts in the respective document. The relevant measure devised contains two parts, where the first part is based on frequent keywords and the second part is based on the remaining keywords of the documents.

1. Frequent keyword-based similarity: The motivation behind this approach is that the most significant words are likely to be referred repeatedly, or, at least, more frequently than unimportant words. In practice, the words that are frequently occurring in a document have more expressive power in the file. Based on this, we have designed a frequent keyword-based similarity measure that gives more importance to the frequent keywords rather than infrequent keyword.
2. Keyword-based similarity: The relevant measure is not a best measure if it is only based on frequent keywords to categorize a web page. To overcome such a situation, we also incorporate keywords other than the frequent words for finding their suitable category. The importance of this keyword-based similarity measure is relatively less compared with the frequency based similarity measure.

Map reduce Programming Models

The Map Reduce programming model is usually used to develop a highly parallel application that process and generate a large amount of data. Map Reduce provides high scalability and reliability because of the

division of the work into smaller units. The data are mainly given to a master key pair value, which is responsible for managing the execution of applications in the cluster. After submitting a job, the master initializes the desired number of smaller tasks or units of work, and puts them to run on worker key pair values. First, during the map phase, key pair values read and apply the map function to a subset of the input data. The map's partial output is stored locally on each key pair value, and served to worker key pair values executing the reduce function.

Map Reduce is extremely appropriate for huge data searching and processing operations. For traditional clusters, the model has shown excellent I/O features, which is apparent from its successful application in large-scale search applications by Google. Data in the Map Reduce framework are usually delineated in the form of key-value pairs <key, value>. The primary step of the computation is the map function where the framework reads input data and optionally changes it into proper key-value pairs. The second step which is the map phase, where on each pair <k, v> a function g which returns a multiset of new key-value pairs is applied. The function is expressed as below,

$$g(< k, v >) = \{< k_1, v_1 >, < k_2, v_2 >, \cdots, < k_n, v_n >\} \quad (1)$$

Where k is the key and v is the value. In the reduce phase, all pairs that are generated in the preceding step are grouped according to their keys and their values are reduced using a function which is given by,

$$h(\{< k_1, v_1 >, < k_2, v_2 >, \cdots, < k_n, v_n >\}) = < k, v > \quad (2)$$

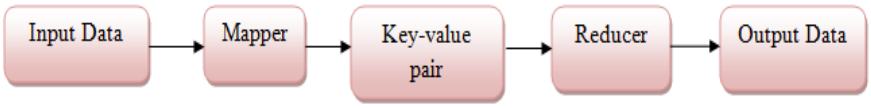


Fig. 1: Map-Reducing Process.

The input dataset which is a document is first applied to the mapper, which generates the key value pair. Here the key is the document name and value is the document content. The same words and the number are grouped as a key-value pair and this is given to the reduce function. The reduce function, then obtains all the pairs of word with the same key and value and counts the number of pairs in the document and then reduces the count by considering the same key-value pair as a single one. Once the map reduces step is over next the data are stored in HDFS. When the program calls MapReduce function, the following sequence of actions occurs sequentially.

The MapReduce function first splits the input files into a number of sections. Then it starts producing a number of copies of the program on cluster of key pair values. Among this, one is the master, which assigns the work to the remaining key pair value which is the workers. There are P map task and Q reduce task. The master splits the task to appropriate workers. The worker who is assigned each map task will read the content of the corresponding split input and generates the key/value pair for that input data and passes each to the Map function. Here we can optimize that key/pair value by employing GA. This can be described in the next section. The intermediate key/value pair produced by the Map function is then given to the memory for storage. The location is then passed to the master key pair value which then forwards this location to reduce workers. When the reduce workers get the location of the intermediate data output, it reads the entire intermediate data and based on the keys they are sorted such that data with the same key are collected into one group. The key and intermediate value set are passed by the workers to the Reduce function. The reduce function produces the final output of the Map Reduce model. This process continues until the entire map and reduce tasks are completed. After the completion of the process; the outputs are stored in the HDFS. The MapReduce program provides the way for better processing as it provides more resources to be available for storage.

The various drawbacks that exist in the related works are being sorted out by our proposed method and based on our proposed algorithm better resource problem solving strategies are obtained. The advantages of our proposed algorithm include,

- The Data involved in one Map/Reduce job is lesser. Hence lesser load on the buffers and better I/O.
- Datasets of any size can be utilized provided there is space available on the HDFS.
- Any number of datasets can be used given enough space on the HDFS.
- MapReduce may be easier for users to adopt for simple or one-time processing tasks.

EVOLUTIONARY ALGORITHMS

Camel In 1975, John Holland introduced a new way to solve problems with computers: Genetic algorithms (GAs) [22]. The GA is a heuristic search technique that simulates the processes of natural selection and evolution. GAs tends to find good and novel solutions to hard problems in a reasonable amount of time. The selection, crossover, mutation and fitness functions are discussed below.

Genetic Algorithm

Genetic algorithms are adaptive methods which may be used to solve search and optimization problems. They are based on the genetic processes of biological organisms. According to the principles of natural selection and survival of the fittest; natural populations are evolved in many generations. By simulating this process, genetic algorithms are able to use solutions to real world problems, if they have been suitably encoded. The genetic algorithm usually works as given below,

Generation and Selection

In genetic algorithm, a population is created with a group of individuals or chromosomes.

$$D = \{D_0, D_2, D_3, \cdots, D_{n-1}\} \quad , 0 \le j \le N_p - 1 \qquad 0 \le k \le n-1 \quad (3)$$

Here D represents the database with collection of key pair values. Among them we have to select the optimal one. The selection process decides which of the chromosomes from the population will be selected for crossover to create new chromosomes. This new chromosome will now include with the population to determine the next selection. The individuals with more fitness value will be selected. The selection is based on the fitness of the individuals.

Pseudo code for chromosome generation:

```
function Chrom = Generate_Chrom(Psize)
 Chrom =[];
 % Psize = 10;
 for t1 = 1 : 10
     Chrom =[Chrom ; randi([1,Psize])];
 end
```

Crossover

After selecting the individuals the next step is the crossover where two parents are made to mate each other. In crossover there are various types in which two point crossovers is more commonly used method. Here the offspring is produced by selecting two crossover point and the genes in between these two points are interchanged from the parents to form new offspring's.

Pseudo code for Crossover:

```
                Function C = CrossOver (C, Psize)
for t = 1 : size(C,1)
                %     tmp1 = C (t,1);
                %     tmp2 = C (t+1,1);
                %     C (t+1,1) =  randi([1,Psize]);
                    C (t,1) =  randi([1,Psize]);
                end
```

Mutation

After crossover, new set of populations are produced. In order to provide individuality to each chromosome, mutation operation is performed where we replace any value with a new value to form a new individual. Among different types of crossovers, the two point crossover is selected with the crossover rate of CR.

In the two point crossover, two points are selected on the parent chromosomes using the eqns. The genes in between the two point's c1 and c2 are interchanged between the parent chromosomes and so Np/2 children chromosomes are obtained. The crossover point's c1 and c2 are determined as follows.

$$c_1 = \frac{\left|D_k^{(j)}\right|}{3} \quad (4)$$

$$c_2 = c_1 + \frac{\left|D_k^{(j)}\right|}{2} \quad (5)$$

Pseudo code for Mutation:

```
                    Function M = Mut(M,Psize)
                r = randi([1 2],1);
for t = 1 : r
                r1 = randi([1 size(M,2)],1) ;
               M (t,r1) =randi([1,Psize]);
                    End
```

Fitness function

After mutation the fitness of each individual is founded and the individual with high Fitness values are selected as the final solution. In our proposed method, the genetic algorithm is used to find out the suitable key pair value for transmission. Here the key pair value is represented as the bit of chromosome. The [Fig. 2] shows flow diagram of genetic algorithm for our proposed method. Initially generate 'N' number of chromosomes (key pair values). Next fitness for each of the individuals is calculated. In our proposed method the fitness value is calculated in relation to the distance between the key values. The key pair values with less distance are selected. Here a Euclidean distance from each key pair value to the next closest key pair values is calculated. It is given by the expression,

$$Fitness(F) = \frac{1}{\min}\left(E_d(m,n)\right) \quad (6)$$

Where m and n are the source and the destination pair values respectively.

Once fitness of the key pair values is calculated we proceed to next step in the genetic algorithm. Next is the selection process where the two chromosomes for crossover and mutation are selected. This selection is based on the fitness of the chromosome. More fit the chromosomes are more the chance for selection. There are various methods of selection in genetic algorithm. In our proposed method we use the roulette wheel selection method. The roulette wheel selection is used for selecting potentially useful solutions for recombination. In roulette wheel selection the chromosome with higher fitness value when compared to others are selected to form the new offspring's. The probability of selecting the 'k'th value is given by,

$$P_k = \frac{F_k}{\sum_{t=1}^{N} F_t} \quad (7)$$

Where the total number of values, P_k is the probability of the 'k'th key pair value and F_k is the fitness of the 'k'th value.

After selecting the solutions crossover and mutation are performed. In the proposed method we utilized two-point crossover method. Two points are selected in the parent chromosomes as R1 and R2 and the genes in between these two points are interchanged to form new offspring's. After the crossover operation, mutation is applied to the newly formed offspring's in order to make each individual independent of the other. After the mutation operation finally the fitness value of the newly formed individuals is calculated. By calculating the fitness value for each individual or key value pair, the values are analyzed and the key pair values with higher fitness values are selected as the most suitable key pair value for transmission. After selecting the optimal value, the map reducer process will be proceeded to get maximum resources.

SIMULATION RESULTS

A series of simulations are carried to evaluate the performance of the proposed GA to solve the resource allocation problems. In our experiment we have utilized the Hadoop map reduce in order to reduce the number of repetitions of the words for more than one time, so the resource memory can be utilized for more storage process. A large set of the databases is collected with different words and the key value pair for each line has to be found out. Here genetic algorithm is used to select the key value pair based on the fitness of the key value pair selected. We have used certain database where different key value pairs are generated based on the number of repetitions of the words. The above key value pair is selected using the map reduce programming and these are based on the availability of the words in the database. Next the selection of the key value is done with the help of the genetic algorithm. The process of selecting the key value pair depend on the fitness of the selected pair. Here the fitness of the key value calculates base the number of times the key word appears in a specific line of the document database. In our experiment we have taken the above keywords as the key value pair and they showed to be fit for selection thus reducing the extra time required in the process of analyzing the document thus saving the resources in a better way when compared to other methods.

It has been seen that our proposed method of map reduce using the genetic algorithm has shown a remarkable reduction in the execution time when compared to other existing method. The map reduces without using the genetic algorithm has shown that it takes more execution time when compared with our proposed method of map reduce using the genetic algorithm. Based on the parameters experimentally determined, the effect of data size, number of clusters, degree of cluster distinctness, degree of cluster asymmetry and level of data randomness at the execution time and clustering quality of the four fast clustering algorithms were evaluated empirically. A discrete event simulator has been developed to evaluate the transition time of the reconfiguration plan obtained by the reduced map approach with GA and the performance of the proposed algorithm in consolidating size of resources.

Transition Time of the Reconfiguration Plan

In the experiments, three types of resources are simulated: CPU, memory and I/O, and three types of VMs are created: CPU-intensive, Memory-intensive, and I/O intensive VMs. A Virtual cluster (VC) consists of the same type of VMs. For the CPU-intensive VMs, the required CPU utilization is selected from the range of [30%, 60%], while their memory and I/O utilization are selected from the range of [1%, 15%].

IIOAB JOURNAL
ISSN: 0976-3104

The VMs are first generated in physical nodes according to the above method. A node is not fully utilized and will have a certain level of spare resource capacity. The service rate of requests of each VM is calculated using the performance model. The workload manager is used in the experiments. The arrival rate of the incoming requests for each VC is determined so that the VCs' QoS can be satisfied. The average execution time for each type of requests is set to be 5 seconds, and the QoS of each VC is defined as 90% of the requests' response time is no longer than 10 seconds. A VC's workload manager (LM) dispatches the requests to VMs, and therefore the request arrival rate for each VM can be determined. Then the developed GA is applied to consolidate VMs so that the spare resource capacity in nodes can converge to a smaller number of nodes. After the GA obtains the optimized system state, the reconfiguration plan is constructed to transfer the Cloud from the current state to the optimized one.

The average time for deleting and creating a VM is 20 and 14 seconds, respectively. The migration time depends on the size of VM image and the number of active VMs in the mapping nodes. The migration time in our experiments is in the range of 10 to 32 seconds.

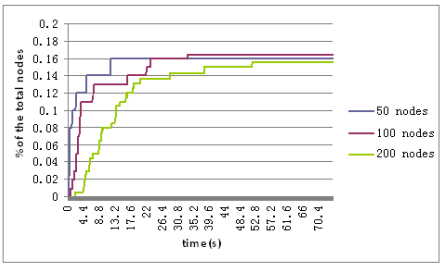


Fig. 2: The quantity of nodes saved as the GA progresses.

[Fig. 2] shows the number of nodes saved as the GA progresses. In the experiments in [Fig. 2], the number of nodes varies from 50 to 200. The experiments aim to investigate the time that the GA needs to find an optimized system state, and also investigate how many nodes the GA can save by converging spare resource capacities. The free capacity of each type of resource in the nodes is selected randomly from the range [10%, 30%] with the average of 20%. The number of the VMs in a physical node is 3. The number of the VCs in the system is 30. As can be observed from [Fig. 2], the percentage of nodes saved increases as the GA runs for longer, as to be expected. Further observations show that under all three cases, the number of nodes saved increases sharply after the GA starts running. It suggests the GA implemented in this paper is very effective in evolving optimized states. When the GA runs for longer, the increasing trend tides off. This is because that the VM-to node mapping and resource allocations calculated by the GA approaches the optimal solutions. Moreover, by observing the difference of the curve trends under a different number of nodes, it can be seen that as the number of nodes increases, it takes the proposed GA longer to approach the optimized state.

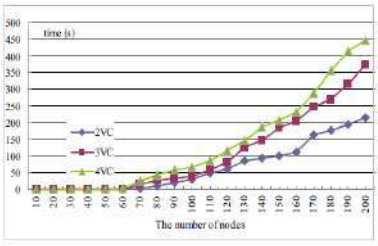


Fig. 3: The execution time of the proposed reconfigure GA algorithm for different number of nodes and VCs.

[Fig. 3] shows the time it takes for the proposed algorithm approach to find the optimal reconfiguration plan under a different number of nodes and different number of VCs. The optimized system states are computed by the GA. The average spare capacity in nodes is 15%. It can be seen from this figure that the time increases as the number of nodes increases and also as the number of VCs increases. When the number of nodes is 200 and the number of VCs is 4, the time is 450 seconds, which is unbearable in real systems. That is why a proposed approach is necessary to quickly find the sub-optimal reconfiguration plan for the large scale of systems. A GA is developed to compute the optimized system state and consolidate resources. The modified map reduce model is then developed to transfer the Cloud from the current state to the optimized one computed by the GA.

Effect of Resource Data Size

The parameters for the synthetic data generation program are summarized in [Table 1]. The default values of these parameters for synthetic resource data sets were n=3000, k=20, t=0.2, a=1, and r=1%. Depending on the type of experiment conducted, the respective parameter was varied, while the rest of parameters adopted their default values. For example, to evaluate the data size effect on the target

clustering algorithms, the parameter n took values from 500 to 7000, while default values were used for the rest of the parameters.

Table 1: Parameters and default values for synthetic data generation

Symbols	Meanings	Defaults
n	Number of resources in a cloud server	3000
k	Number of clusters	20
t	Degree of cluster distinctness	0.2
a	Degree of cluster asymmetry	1
r	Level of data randomness	1%

Synthetic data sets were generated for various data sizes, ranging from 500 to 7,000 (i.e., n=500, 1,000, 2,000, ..., 7,000). Remaining parameters received their default values, as defined in [Table 1]. [Fig. 4] shows the performance of the target clustering algorithms as a function of the data size.

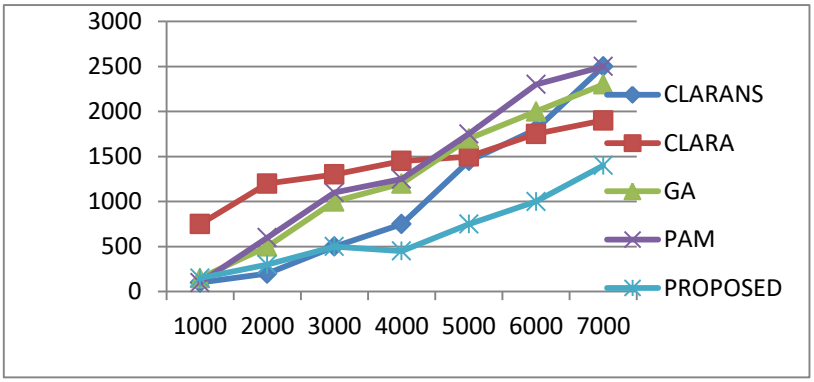


Fig. 4: Effect of Resource data size.

As shown in [Fig. 4], PAM and CLARA slightly outperformed the others in terms of clustering quality when given only a small data size (i.e., fewer than 1,000).When the data size was increased, the clustering quality of the proposed algorithm degraded as compared to that of others. In terms of execution time, our proposed algorithm is more efficient than CLARA, CLARANS, GA and PAM when the data size was more than 3000 sec. As the data size increased, CLARA increased its execution time increased, but our proposed algorithm is a very good solution for the large resource data size communication between cloud servers.

Table 2: Comparison of execution time of the proposed method using genetic algorithm and existing methods

Resource Data size	Execution time(sec)				
	CLARANS	CLARA	GA	PAM	Proposed Algorithm
1000	100	750	150	100	150
2000	200	1200	500	600	300
3000	500	1300	1000	1100	500
4000	750	1450	1200	1250	450
5000	1450	1500	1700	1750	750
6000	1800	1750	2000	2300	1000
7000	2500	1900	2300	2500	1400

In the [Table 2], the execution time for the proposed and existing methods is shown. As it indicates, the execution time for the proposed method improved over time and this proved to be more effective when compared with the existing methods. Based on the execution time for different numbers of nodes the graph is plotted for comparing the performance of different algorithms which are used in map reduce programming. As shown in [Fig. 4], the execution time for our proposed method of map reducing using the genetic algorithm proved to be more reliable when compared to the other algorithms.

CONCLUSION

The Genetic algorithm, Partioning around Medoids, Clustering Large Applications, and Clustering large applications based on randomized search is discussed to solve resource problems in the cloud. Here we

employed a Map Reduce programming model for distributed parallel computing and execution of virtualization process which is used to detect non-sufficient reductions in the execution time and to detect the decrease in the computing time. The proposed method proved to be efficient, one when considering the outcome compared to other algorithms. The graph shows the execution time is reduced to a large instant when compared to the existing method. The simulation results confirm to be an efficient method in reducing the resource problems that occur in the cloud computing. Further enhancement can be done to reduce execution time for larger networks with the number of node and to testify the reliability and securability of the resource allocation where the execution time decreases considerably that enhances the horizon for further study.

CONFLICT OF INTEREST

There is no conflict of interest regarding this manuscript

ACKNOWLEDGEMENTS

There is no acknowledgement regarding this manuscript.

FINANCIAL DISCLOSURE

There is no financial support for this manuscript.

AUTHOR CONTRIBUTION

I Arivudainambi contribute to development and growing popularity of cloud computing indicates the evolution in the way IT infrastructure and services are distributed and expended. This increased use of cloud computing resulted in resource problems which have to be solved for better usage of the clouds.

I Dhanya contribute to present an efficient method for solving resource problems in the cloud using modified map reducing algorithm. Here we employed a MapReduce programming model with GA for distributed parallel computing and execution on a virtualization process which is used to detect non-sufficient reductions in the execution time and to detect the decrease in the computing time.

REFERENCES

[1] Arivudainambi D, Dhanya D. [2013] Improving MapReduce Performance in Cloud Using Genetic Algorithm, 1060-1067, Conference Proceedings, International Conference on Mathematics and Computer Engineering.

[2] Chaisiri S, Lee BS, Niyato D. [2009] Optimal Virtual Machine Placement across Multiple Cloud Providers, IEEE Asia-Pacific Services Computing Conference (IEEE-APSCC). 103-110.

[3] Lucas-Simarro JL, Vozmediano RM, Montero RS, Llorente IM. [2013] Scheduling strategies for optimal service deployment across multiple clouds, Future Generation Computer Systems. 29:1431-1441.

[4] Tordsson J, Montero RS, Vozmediano RM, Llorente IM. [2012] Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers, Future Generation Computer Systems. 28:358-367.

[5] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. [2009] Cloud computing and emerging IT Platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems. 25: 599–616.

[6] Mian R, Martin P, Vazquez-Poletti JL, [2013] Provisioning data analytic workloads in a cloud, Future Generation Computer Systems , 29:1452-1458.

[7] Zhu Y, Hu H, Ahn GJ, Yau SS. [2012] Efficient audit service outsourcing for data integrity in clouds, The Journal of Systems and Software. 85:1083– 1095.

[8] Srirama SN, Jakovits P, Vainikko E. [2012] Adapting scientific computing problems to clouds using MapReduce, Future Generation Computer Systems. 28:184–192.

[9] Liu Q, Wang G, Wu J. [2012] Secure and privacy preserving keyword searching for cloud storage services, Journal of Network and Computer Applications. 35:927–933.

[10] Pallickara SL, Pierce M, Dong Q, Kong CH. [2009] Enabling Large Scale Scientific Computations for Expressed Sequence Tag Sequencing over Grid and Cloud Computing Clusters, In Proceedings of the 8th International Conference On Parallel Processing And Applied Mathematics.

[11] Seinstra FJ, Maassen J, Van Nieuwpoort RV, Drost N, Kessel TV, Werkhoven BV, Urbani J, Jacobs C, Kielmann T, Bal HE. [2011] Jungle Computing: Distributed Supercomputing beyond Clusters, Grids, and Clouds, Computer Communications and Networks. 167-197.

[12] Kim M, Lee H, Cui Y. [2011] Performance Evaluation of Image Conversion Module Based on Map Reduce for Transcoding and Transmoding in SMCCSE, IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing. 396-403.

[13] GonzaLez-Ve´Lez H, Kontagora M. [2011] Performance Evaluation Of Mapreduce Using Full Virtualisation On A Departmental Cloud, International Journal of applied mathematics and computer science. 21(2):275–284.

[14] Grossman R, Gu Y. [2008] Data mining using high performance data clouds: experimental studies using sector and sphere, In Proceeding of the 14th ACM SIGKDD International conference on Knowledge discovery and data mining, New York. 920-927.

[15] Papadimitriou S, Sun J. [2008] DisCo: Distributed Co-clustering with Map-Reduce: A Case Study towards Petabyte-Scale End-to-End Mining, In Proceedings of the 8th IEEE International Conference on Data mining, Hawthorne. 512-521.

[16] Zhang C, Sterck HD. [2009] CloudWF: A Computational Workflow System for Clouds Based on Hadoop, Cloud computing. 5931:393-404.

[17] Jin C, Buyya R. [2009] Map Reduce Programming Model for .NET-based Distributed Computing, In Proceedings of the 15th International Euro-Par Conference on Parallel Processing, Berlin. 417-428.

[18] Sehgal S, Erdelyi M, Merzky A, Jha S, [2011] Understanding Application-Level Interoperability: Scaling-Out Map Reduce Over High-Performance Grids and Clouds , Future Generation Computer Systems, 27: 590-599.

[19] Marozzo F, Talia D, Trunfio P. [2008] Adapting Map Reduce for Dynamic Environments Using a Peer-to-Peer Model , In Proceedings of the 1st Workshop on Cloud Computing and its Applications.

[20] Tian F, Chen K. [2011] Towards Optimal Resource Provisioning for Running Map Reduce Programs in Public Clouds, In Proceedings of the IEEE International Conference on Cloud Computing (CLOUD). 155-162.

[21] Yang X, Zola J, Aluru S. [2011] Parallel Metagenomic Sequence Clustering via Sketching and Maximal Quasi-clique Enumeration on Map-reduce Clouds, In Proceedings of the 2011 IEEE International Conference

on Parallel & Distributed Processing Symposium (IPDPS). 1223-1233.
[22] Husain M, McGlothlin J, Masud MM, Khan L, Thuraisingham B. [2011] Heuristics Based Query Processing for Large RDF Graphs Using Cloud Computing, IEEE Transactions on Knowledge and Data Engineering. 23(9):1312-1327.
[23] Holland JH. [1975] Adaptation in natural and artificial systems: an introductory analysis with applications to biology, Control and artificial intelligence. Ann Arbor: University of Michigan Press. 175-177.
[24] Patel K, Annavaram M. [2013] NFRA: Generalized Network Flow Based Resource Allocation for Hosting Centers, IEEE Transactions On Computers. 62(9):1772-1785.
[25] Meri K, Arenas MG, Mora AM, Merelo JJ, Castillo PA, Garcia-Sanchez P, Laredo JLJ. [2012] Cloud-based evolutionary algorithms: An algorithmic study, Natural Computing-Springer. 12(2):135-147.
[26] Zhu K, Song H, Liu L, Gao J, Cheng G. [2011] Hybrid Genetic Algorithm for Cloud Computing Applications, IEEE Asia-Pacific Services Computing Conference. 182-187.
[27] Kuribayashi SI. [2011] Optimal Joint Multiple Resource Allocation Method for Cloud Computing Environments, International Journal of Research and Reviews in Computer Science(IJRRCS). 2(1).
[28] Yu F, Wan YW, Tsaih RH, [2013] Quantitative Analysis of Cloud-based Streaming Services, IEEE International conference on service computing (SCC). 216-223.
[29] Salvadores M, Zuo L, Imtiaz SMH, Darlington J, Gibbins N, Shadbolt NR, Dobree J. [2010] Market Blended Insight: Modeling Propensity to Buy with the Semantic Web, Lecture Notes in Computer Science. 5318:777-789.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

CORPUS BASED SENTIMENT CLASSIFICATION OF TAMIL MOVIE TWEETS USING SYNTACTIC PATTERNS

Nadana Ravishankar[1*], Shriram Raghunathan[2]

[1,2]Department of Computer Science and Engineering, B.S. Abdur Rahman University, Chennai, INDIA

ABSTRACT

In a practical scenario, word of mouth is the traditional way for movie recommendation. Reviews provided by friends and/or relatives about a movie or product are used whether to see the movie or not. In this research, we address the genre classification of Tamil movie tweets based on sentiments expressed by users in addition to opinion mining. We built our own dataset contains of 7842 tweets expressing sentiments regarding Tamil movies using Tamilagarathi. Then, we evaluate data collection techniques and different sentiment categorizing approaches to predictive algorithms that are used as tool to identify the sentiment categories stated in micro-blogging site, known as Twitter. We present that applicability of the predictive model on movies with popularity levels by analyzing the user tweets and discuss its empirical evaluation methods using accuracy as a metric. Finally, we compare the results with existing baseline models like TF-IDF about a Tamil movie in concern to recommend people to choose whether to watch the movie or don't, producing significant results.

KEY WORDS
Sentiment Analysis, Tamil tweets, NLP, Categorization, Opinion mining

Received: 26 March 2017
Accepted: 15 June 2017
Published: 10 July 2017

***Corresponding Author**
Email:
nadanaravishankar@gmail.com
Tel.: +91-44 27465315

INTRODUCTION

The popularity of social media has rapidly increased in today's digital world. People actively participate in online activities and surveys by voluntarily posting messages in their mother languages and emoticons for example product reviews, movie reviews, political issues, etc. Twitter is a famous social media service that supports users to send messages up to 140 characters in length, called tweets. As of January 2016, Twitter had 313 million monthly active users and they are posting an average of 550 million of Tweets per day[1]. Most of the Sentiment Analysis (SA) aims to find the polarity (positive, neutral and negative) of a keyword or text, or to make sentence or text level classification based on their polarity score [1].

The proliferation of Twitter has generated a source of extremely large data and most of the information is publicly accessible to all users. Twitter users express their opinions about a wide range of issues like movie, politics, technology, books, religion, food, sports etc. Mining this volume of sentiments provides information for understanding collective human behavior and it is of valuable commercial interest to a particular movie, product or service. Most of the aspects shared by all these opinions are the subjectivity, since the sentiments expressed by the users about a product or service is not biased. Sentiment analysis of data collected from Twitter users can be able to extract the general emotion of users in relation to a range of topics. These sentiments can be very useful for companies to decide about their competitors and user feelings about their product and also useful for individuals to decide whether to purchase the product or not.

Patterns mining are the new field where the lack of supervision and local results intersects, which holds a great promise for future: with large amounts of data given and no supervision, PM can find interesting relationships in the data (corpus). Those patterns can in turn be used to any domains. Our main objective is to propose a domain-specific model for finding sentiment category based on syntactic models like TF-IDF and tweet weightage. Our main focus is on developing a model for Tamil movie domain since, as several research review on sentiment analysis demonstrate [2-3], it is a very difficult to find the user sentiments in natural language and adapt the user content in sentiment analysis.

The remaining sections of the article are organized as follows. Section 2 discusses some related works. Section 3 presents the different approaches we use different categorization models for classification of user sentiments and in section 4, results regarding categories and comparison regarding movies used as domain. Finally, in Section 5 we present our conclusions and outline future works.

RELATED WORK

The field of text categorization was introduced long time ago [4], however categorization based on sentiment was introduced more recently in [5-7]. The standard approach for text representation has been the bag-of-words (BOW) method. According to the BOW model, the document is represented as a vector of words in Euclidian space where each word is independent from others. This bag of individual words is commonly called a collection of unigrams. The BOW is easy to understand and allows achieving high performance.

People are linking each other with the support of internet through online conservation forums, blog post and much more [8]. Neethu and Rajasree [9] have stated that people view the ratings or reviews of movies before watching that movie in theatres. In [10] author has mentioned that most analysis work of sentiment has been performed on review (movie) sites. Review sites offers with opinions of movies or products thus limiting the application domain to only business.

Another research in [11], mentioned about the prediction model for sentiment analysis of tweets of movie success. Author has used the sentiment analysis tools with the Naïve Bayes classifier using the NLTK tool kit. He has used the tweets statistics to classify the hit/flop/ average movies. Thigale et al [12] studied about the box office success of Hollywood movies using the publicity analysis of twitter data prediction. Regression methods are act as the effective tool of forecasting and predicting the revenue of the particular thing using the social media. Authors of recent research [13], studied about the calculating sentiment scores for every released movie from twitter user data using the python module Sci-Kit learning tool has been selected as the classification tool for the classification of machine learning based experiments.

The goal of this research is to mine all the tweets and present it in a form to the user to help take their decision. In this research, we consider five pre-defined categories of genres for Tamil movies: action, love/sentiment, commercial, family and comedy. We also used different tokenization and pre-processing approaches to predict the best combination of substitutes that aims to better performance in the domain of movies categorization.

PROPOSED METHODS

[Fig. 1] shows the framework for genre classification of Tamil tweets. We collect tweets and stored them in a database and build our own corpus. After formatting the dataset, apply different sentiment categorization algorithms to find the user sentiments about a movie and present it to the user to decide whether to watch it or not.

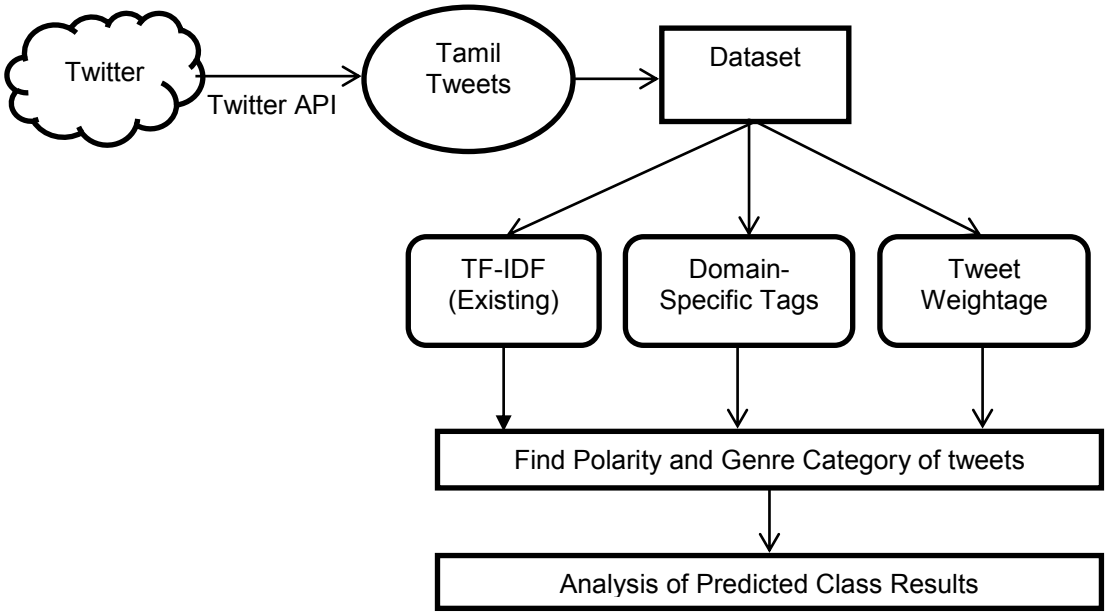


Fig. 1: Syntactic based sentiment analysis model for classification of Tamil movie tweets

Data collection and pre-processing

We have collected data from Twitter by using #Hashtag followed by the movie name such as # வீரம் (Veeram), # கபாலி (Kabali) etc. We use the API provided by Tamil Agarathy (http://agarathi.com/api/dictionary#) to derive the sentiment related information from the dataset. The API provides Tamil dictionary that has 1 lakh words and its syntactic meaning. The words are grouped according to their lexical and conceptual relations. However the dictionary doesn't provide any sentiment related information and we leverage this dictionary to derive the sentiments. As of July 2016, we didn't find any data for Tamil movie from the literature, we have developed our own corpus for around 100 Tamil movies and around 7,000 tweets have been stored in dataset for our experimental purpose.

Tokenization process was followed after the data pre-processing task. The steps involved in pre-processing task is: the removal of any external links (URL's) and retweets and removal of any characters that repeat more than one such as **ஆம்ம்** (Yes) to **ஆம்**. This process is shown in [Table 1].

Table 1: The process of tokenization

Before	After Tokenization
பார்த்த பாலா படங்களில் ரசித்த உணர்வுப்பூரணமான கதை #பரதேசி Partha bala padangalil rasitha unarvuppuramana story #paradesi	பார்த்த partha (1) பாலா bala (2) படங்களில் padangalil (3) ரசித்த rasitha (4) உணர்வுப்பூரணமான unarvuppuranamana (5) கதை Story (6)

Term Frequency- Inverse Document Frequency (TF-IDF) ranking

TF-IDF is one of the simplest approaches for text classification [4]. TF-IDF works fine in documents classification, like news articles or reviews. However, we found from the literature that TF-IDF does not classify tweets well as tweets are short in length, don't follow grammar style and generally words repeat rarely [13- 14]. We choose TF-IDF as our baseline since it provides the importance of a word in a data set. Tweets contain set of words, so the most frequent words should correspond to the topics, obtaining the most relevant words. For each movie, the top n TF-IDF keyword values are selected to categorize the tweets in a dataset. In this research work, we have a set of genre categories for Tamil movie domain.

Consider a movie m_i which is associated with a set of tweets $\{t_1, t_2 \ldots t_n\}$.
Each tweet is made up of a set of terms and so each movie can be characterized as a sequence of words $w_1, w_2 \ldots w_k$. The $tf(w_j, m_i)$ and $idf(w_j, m)$ values are calculated as follows.

$$tf(w_j, m_i) = \frac{\text{frequency of occurrence of } w_j}{m_i} \quad \ldots\ldots\ldots (1)$$

$$idf(w_j, m) = \log\left(\frac{\text{Total tweets in overall movie (m)}}{\text{frequency of occurance of } w_j \text{ in no of tweets}}\right) \quad \ldots\ldots\ldots (2)$$

Consider a **வீரம்** movie containing 305 tweets wherein the word **வசூல்** appears 20 times.

$$tf_{\text{வசூல்}} = \frac{20}{305} = 0.0656 \qquad idf_{\text{வசூல்}} = \log\frac{305}{18} = 1.229$$

$$tfidf_{\text{வசூல்}} = tf * idf = 0.0806$$

Likewise we have calculated the TF-IDF score of all genres corresponding to the domain and this score can be used to identify the important keywords. The most related words to the genres are also mapped to the appropriate category using Tamil dictionary. For the genre category of **சண்டை** (sandai), the closely related words like **அதிரடி** (athiradi), **போர்** (por), **மோதல்** (mothal) etc., are mapped to the same category.

Classification of Tamil movie tweets using TF-IDF provides a baseline for our proposed approach; we find that no other research work has been carried out for genre classification of Tamil tweets. The accuracy result of TF-IDF is shown in [Table2]. The accuracy metric is calculated as the correct predicted categories of tweets by total number of tweets.

The accuracy of the existing model is challenging as compared to English tweets as in [15-16]. The reason is the Tamil dictionary is developed from the lexicon based word formation rules and not focused for user tweets. We also find the polarity of the tweets in addition to genre classification of Tamil tweets using the dataset as mentioned in [17].

Domain-specific Tags (DST)

TF-IDF can only analyse the presence of a word and its syntactical words in the dataset. We can extend it by incorporating domain-specific tags to cover all the words in dataset. Twitter users don't follow linguistic rules and pure lexicons in their tweets. The aim of this method is to consider all the slang words and non-

English words and match those words to Tamil dictionary to enhance the accuracy of the classification model.

Table 2: Accuracy of Existing model (TF-IDF) for the movie **வீரம்**

Categories	TF-IDF Score	Accuracy
$tfidf_{சண்டை}$	6.98	**27.13**
$tfidf_{வதூல்}$	8.06	
$tfidf_{காதல்}$	2.92	
$tfidf_{காமெடி}$	5.36	
$tfidf_{சென்டிமென்ட்}$	6.58	

We have manually analyzed the tweets and developed a dataset in addition to Tamil dictionary to map all slang words and non-English words to find the sentiments within the tweets. Four groups have participated and annotated these dataset. We have built a dataset for DST for 1000 most occurring words in tweets. The DST for the sample words are given in [Table3].

Table 3: Sample dataset model used in DST

Words	Meta data	Polarity	Primary category	Secondary category
சண்ட	சண்டை, ஆக்சன், பைட்டு, அதிரடி	-	சண்டை	வதூல்
மொக்க	மொக்கை, கடி, அறுவை, வெறுப்பு	எதிர்மறை	-	-
செம	அற்புதம், செம்ம மிகப்பெரிய	நேர்மறை	வதூல்	-
காமெடி	நகைச்சுவை, ஹாஷ்யம், சிரிப்பு, புன்னகை	நேர்மறை	காமெடி	-

The aim of this approach is to check whether the incorporation of tweets focused words increase the accuracy of the model or not. However we have covered all the words in the tweets dataset, accuracy is slightly increased as shown in [Table 4].

Table 4: Performance analysis of TF-IDF and DST

Movie	Accuracy	
	TF-IDF only (Existing)	TF-IDF + DST (Our contribution)
வீரம்	27.13	36.42

TWEET WEIGHTAGE MODEL

TF-IDF and DST can analyze the significance of a single keyword and its related words in the dataset but doesn't consider the importance of co-occurring words. For tweet weight approach, we invoked the method used in [18] and applied it on sentences to find the sentiment category rather than a single word. Based on literature, we formed a hypothesis that a set of patterns have been built for the domain of interest and most of the tweet length is less than 8 words in relation to Tamil movie domain. When the tweet length is more, the complexity of the words in the tweets is difficult to model. Hence lesser weightage can be given for tweets with more number of words. We find that our contribution is the evaluation of sentence-level sentiment categories after pruning, as explained in the next steps.

IIOAB JOURNAL
ISSN: 0976-3104

- Let us take two groups S_1 and S_2
- If tweet length is less than 8, then it is added to group S_1, otherwise S_2
- Assign weight values for S_1 and S_2
- Calculate the polarity and category of tweets using Tamil dictionary.
- We have experimented with different weight values for S_1 and S_2 and compared the accuracy values.
- Finally we choose the optimal S_1 and S_2 values as feature values.

Table 5: Performance of tweet weight model for different weight values

Movie Name	S_1	S_2	Accuracy
வீரம்	0.3	0.7	30.64
	0.4	0.6	31.42
	0.5	0.5	33.15
	0.6	0.4	38.97
	0.7	0.3	**40.26**

Results indicate [Table 5] an improvement in classification accuracy when using the proposed tweet weightage algorithm compared to TF-IDF and DST. Tweet weight based model incorporate sentence structure better to identify syntactic information for improving sentiment classification. However, sentence length cut-off used in this approach may result in low accuracy due to loss of important information but it is purely depend on sentiment lexicon of that particular domain of interest.

RESULTS

We use the Python programming language and the Natural Language Toolkit's (NLTK) implementation; algorithms were implemented to determine sentiment found within tweets stored in the dataset. User tweets about a particular movie is categorized into any one of the type as shown in [Table 6].

Table 6: The types of sentiment category in Tamil movies

Category	Description
சண்டை (Sandai)	denotes a tweet belongs to action category
காதல் (Kadhal)	denotes a tweet belongs to Romance category
மசாலா (Masala)	denotes a tweet belongs to commercial category
குடும்பம் (Kudumbam)	denotes a tweet belongs to Family sentiment
காமெடி (Comedy)	denotes a tweet belongs to Comedy category

From the above discussions, we developed a set of syntactical models for classification of a Tamil movie dataset. In this part, we validate the performance (accuracy) of all proposed sentiment categorization models for Tamil movie tweets: given new movie name like **வீரம்**, we aim to find the user sentiments based on the proposed model for given Tamil movie composed from Twitter. We calculate the average accuracy of different models for all the Tamil movies present in dataset.

For each given movie, we search for tweets using the movie name (in Tamil) from the database. We created a dataset for 100 Tamil movies and its related tweets for each movie, which was then, annotated automatically using the keywords and metadata built as defined in the previous sections. Once we select the type of sentiment category model to implement, the category predicted by the sentiment model for each tweet was manually analyzed by the group of domain experts and their scholars in order to compute the average accuracy of each categorizer models.

IIOAB JOURNAL
ISSN: 0976-3104

Fig. 2: Screenshot of sentiment analysis for a Tamil movie Kabali (கபாலி)

Our results for different sentiment categorization approaches are presented in [Table 7]. We can observe that the TF-IDF algorithm had the lowest accuracy of 29.87%. The reason may lay in attribute selection and does not consider the context of the domain needed to extract the original sentiments contain in a tweet. Tweet weightage and DST methods have produced improved results (average) but better than TF-IDF model.

TABLE 7: Overall Accuracy results of sentiment analysis methods

Method	Average Accuracy
TF-IDF ranking	29.87
TF-IDF + DST	35.64
Tweet Weightage	**40.07**

The best syntactic model was tweet weightage model that achieved the accuracy of 40.07%. Comparing our results with the results of the TF-IDF and DST, we can conclude that our proposed algorithm based on the weightage scheme would have produced the best accuracy of 15% higher than other models. We also find the average recall for each Tamil movie and we validate that the best model (according to metrics) were the syntactic based tweet weightage models discussed above.

CONCLUSION AND FUTURE WORK

In this research, we evaluated our proposed syntactic algorithms to build genre classification models of Tamil tweets and compared the average accuracy of algorithms to find sentiment category of user tweets about a set of Tamil movies stored in the dataset and to help users to take decision corresponding to particular movie. The algorithm can be invoked for different domains and systems.

We found that sentiment category models based on tweet weightage obtain better accuracy results than TF-IDF. To summarize, we can conclude that to develop a good sentiment categorizer model in the context of Tamil tweets, depends mainly on the pre-processing approaches used for Tamil tweets and the algorithms used to categorize them. Corpus creation and Tamil dictionary have been used with the limited resource of linguistic knowledge. Though the algorithm is quite difficult for Tamil language, there is a significant change in accuracy observed in our implemented model. However, there is a need for further improvement and lot of research in linguistic models to understand the context of the domain. In the future, we plan to develop rule based sentiment analysis methods for classification of Tamil tweets. Though we considered only domain-specific tweets in the context of Tamil movies, we argue that our approach can be adapted to other domains such as product or service.

CONFLICT OF INTEREST
There is no conflict of interest regarding this manuscript.

ACKNOWLEDGEMENTS
None.

FINANCIAL DISCLOSURE
None.

REFERENCES

[1] Bing Liu (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015: 381 pages.
[2] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. In Comput. Linguist., 37(1): 267–307.
[3] M Dolores Molina-Gonzalez, Eugenio Martinez-Camara, M Teresa Martın-Valdivia, and L Alfonso Urena Lopez. (2015). A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.
[4] Salton, G. and McGill, M. J. (1983). In Introduction to Modern Information Retrieval. McGraw Hill Book Co.
[5] Das, S. and Chen, M. (2001). Yahoo! for amazon: Extracting market sentiment from stock message boards. In Asia Pacific Finance Association Annual Conf. (APFA).

[6] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, Stroudsburg, PA, USA. Association for Computational Linguistics: 79–86.

[7] Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Stroudsburg, PA, USA. Association for Computational Linguistics: 417–424.

[8] Amolik, A., Jivane, N., Bhandari, M., &Venkatesan, M (2015), Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques, International Journal of Engineering and Technology, 7(6): 2038-2043.

[9] Neethu M, S and Rajasree R (2013), 'Sentiment analysis in Twitter using Machine Learning Techniques', Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Trivandrum: 1-6.

[10] Agarwal A, Xie B, Vovsha I, Rambow O and Passonneau R (2011), "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media: 30–38.

[11] Jain V (2013), Prediction of Movie Success Using Sentiment Analysis of Tweets, The International Journal of Soft Computing and Software Engineering, 3(46): 308-313.

[12] Thigale S, Prasad T, Makhija U K and Ravichandran V (2014), Prediction of Box Office Success of Movies Using Hype Analysis of Twitter Data, International Journal of Innovative Engineering and Science, 3(1): 1-6.

[13] Schmidt W and Wubben S (2015), Predicting Ratings of New Movie Release from Twitter Content, Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), Portugal: 122–126,.

[14] Salud M. Jimenez-Zafra, M. Teresa Martın-Valdivia,M. Dolores Molina-Gonzalez and L. Alfonso Urena-Lopez. Domain Adaptation of Polarity Lexicon combining Term Frequency and Bootstrapping, Proceedings of NAACL-HLT 2016, San Diego, California, June 12-17, 2016: 137–146.

[15] Olena Medelyan, Vye Perrone, and Ian H. Witten. (2010). Subject metadata support powered by maui. In Jane Hunter, Carl Lagoze, C. Lee Giles, and Yuan-Fang Li, editors, JCDL, ACM : 407–408..

[16] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V. (2015). SemEval-2015 task 10: sentiment analysis in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Co-located with NAACL, Denver, Colorado, pp. 451–463.

[17] Braja Gopal Patra , Dipankar Das , Amitava Das , and Rajendra Prasath, (2015). Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview. In Proceedings of the Third International Conference on Mining Intelligence and Knowledge Exploration, Hyderabad, India. Springer-Verlag: 650-655.

[18] Gizem Gezici , Berrin Yanikoglu , Dilek Tapucu , Yücel Saygın. (2012). New Features for Sentiment Analysis: Do Sentences Matter? In Proceedings of the International conference on Knowledge discovery and data mining: 783-792.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

COMPARISON OF DIFFERENT BIOMETRIC FEATURES BASED IDENTIFICATION SYSTEM

S. Suganthidevi*, A. Suhasini,

Department of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamilnadu, INDIA

ABSTRACT

In the emerging technology security identification is one of the major challenges in various sectors because it is used to eliminate the unauthorized access while accessing the secure information. During the authentication purpose, various biometric features such as ear, fingerprint, knuckle, voice, iris are used. Among the different biometric features, in this work utilize the finger knuckle and ear features are utilized for performing the identification process. Based on the different features, the biometric authentication system is implemented with the help of the Graph based Geometric Approach and the Principal Curvature approach and Rapid Segment Analysis Feature Transform (RSAFT) method. Thus in this paper compares the different identification method for authenticate the user information. Then the best identification method is analyzed in terms of different metrics such as false acceptance rate, false rejection rate, equal error rate, and accuracy metrics.

KEY WORDS
Biometric features, finger knuckle and ear features,Graph based Geometric Approach and the Principal Curvature approach and Rapid Segment Analysis Feature Transform (RSAFT)

Received: 28 June 2017
Accepted: 11 July 2017
Published: 30 July 2017

***Corresponding Author**
Email:
suganthidevis@gmail.com

INTRODUCTION

In the developing technology security is one of the major issue in various fields such as cloud computing, bank transaction, educational institutions and so on. This security problem is resolved with the help of the biometric system which effectively utilizes the human traits for authenticating the user identities [1]. There are various human traits such as ear, fingerprint, face, lips, retina, signature, DNA, hand geometry and voice features are used for implementing the effective biometric system. Among the different biometric features, in this work utilizes the two important biometric features namely finger knuckle print and ear biometric feature for implementing the authentication system because those features have several advantages when compared to the other biometric features [2]. The finger knuckle print features having several benefits such as it numerous image capturing, contact-less, user acceptance of the outer –palm surface is high, well defined texture feature, easily accessible and stable features because it does not change their characteristics during the personal emotions and behavioral aspects [3]. In addition the ear biometric features having separate advantages namely, ear features are effectively utilized in the forensic science because it does not change its structure and color while developing the human life cycle. More over the ear biometric features does not affected by human personal emotions; in addition the ear images are captured from very long distance without user interference [4]. Due to these advantages on the two biometrics features in this research work utilizes these features for implementing the identification system. These two biometric features are shown in the [Fig. 1].

The captured biometric features are processed by applying the different image processing approaches such as noise removal, segmentation, feature extraction and matching process. These steps are successfully analyzing the each and every pixel present [5] in the image with effective manner. In this work two different optimized image processing techniques such as Graph based Geometric Approach and the Principal Curvature approach and Rapid Segment Analysis Feature Transform (RSAFT) are introduced for processing the images with effective manner. After implementing the biometric system, comparison should be performed in for analyzing the effectiveness of the system in terms of using the performance metrics such as false acceptance rate, false rejection rate, equal error rate, and accuracy metrics. Thus the remaining of this paper organized as follows, Section 2 summarizes the related works for different biometric feature based identification process, Section 3 deals with the detailed biometric system, and Section 4 discusses the results and comparison of the identification system then the Section 5 describes the conclusion.

RELATED WORKS

In this section examines the various author's opinions regarding the biometric feature based identification system. Sayan Maity et al [6] recognize the system by using the 3D image based new data storage and information retrieval technique. This paper proposes that, University of Notre Dame Collection J2 dataset 3D ear images are segmented by using the automatic segmentation process and categorizes those images by shape and surface based features. Then the geometrical features are estimated with the help of shape and surface values and the features are categorizes using the tree based indexing which was done in tree based balance split and tree based unbalanced split. The resultant of the system produces the enhanced recognized system with minimum time.

Vaibhav et al., [7] develops the novel biometric recognition system using the Radon Transform. The captured finger knuckle biometric image considered as the texture image which is preprocessed and the noise has been removed for improving the recognition rate. Then the different direction based features are extracted using the Random transform which is classified by applying the weighted average difference measures. Then the performance of the system is evaluated using the PolyUFKP database knuckle print

images which are analyzed using the 60 different directions ranging from 0-180 degree with the interval of 3 degrees. Then the feature vectors are mapped according to the 256×60 size, which provides the 94.33% recognition rate.

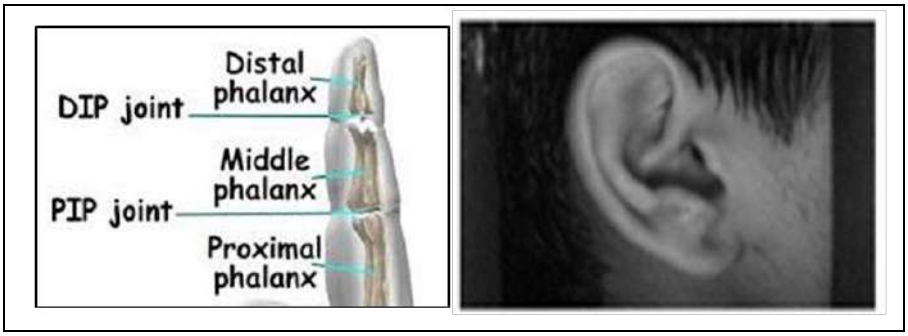


Fig. 1: Sample Biometric Features.

Mahesh Kumar, et al.,[8] proposes a finger print knuckle biometric system for improving the authentication and security to the user personal information's. The captured knuckle print biometric features are preprocessed and the local, global texture features are extracted by utilizing the symmetric discrete orthonormal Stockwell transform.

Meraoumia et al., [9] improves the multimodal biometric system using the palm print and finger knuckle print biometric features. The captured features are processed by using the Phase-Correlation Function (PCF). Then the two biometric features are combined with the help of the matching score level and the performance of the system is evaluated using the recognition rate. Based on the above discussions, the biometric features are processed by applying the different image processing techniques. So, in this paper utilizes the Graph based Geometric Approach and the Principal Curvature approach and Rapid Segment Analysis Feature Transform (RSAFT) method based image processing methods for implementing the identification system which is explained in the following section.

DIFFERENT BIOMETRIC FEATURE BASED IDENTIFICATION SYSTEM

In this section discusses the two different biometric feature based identification system such as Graph based Geometric Approach and the Principal Curvature approach and Rapid Segment Analysis Feature Transform (RSAFT) method implementation process which is explained one by one as follows.

Graph based geometric approach and the Principal curvature approach

First the finger knuckle biometric based identification system is developed by using the Graph based Geometric Approach and the Principal Curvature approach. Initially the finger knuckle image has been captured and the it has been converted into grayscale image[10] by applying the following eqn(1).

$$GS = 0.2989 * Intensity(r) + 0.58701 * Intensity\,(g) + 0.1140 * Intensity\,(b) \quad (1)$$

After converting the image color, noise present in the image has been eliminated by using the non-local median filter. The method effectively analyzes the self-similarity between the pixels in terms of using the intensity value which is measured as follows.

$$v(i) = u(i) + n(i) \quad (2)$$

Where v(i) is defined as the observed value from the given image, u(i) is defined as the "true" value and n(i) is defined as the noise agitation at a pixeli. After that the Gaussian noise has been eliminated by assuming the identical values and the non-local mean[11] value is estimated as follows,

$$NL(V)(p) = \sum_{q \in V} w(p,q)V(q) \quad (3)$$

Where V is defined as the noisy image, and weights w(p,q) meet the subsequent conditions $0 \le w(p,q) \le 1$ and $\sum_q [w(p,q)=1]$. Then the weighted and neighboring values are calculated for determining the noise value which is eliminated successfully. After eliminating the noise in the image contour off the image has been extracted based on the background pixel and object pixel. This pixels are analyzed effectively based on that graph has been constructed which helps to link the contour pixels with effective manner. Then the finger knuckle key point location is identified by applying the principle curvature based detector[12]. The point has been detected in terms of intensity and structure based detector. The principal curvature region is calculated by applying the Hessian Matrix, which is defined as follows,

$$H(x) = \begin{bmatrix} L_{xx}(x) & L_{xy}(x) \\ L_{xy}(x) & L_{yy}(x) \end{bmatrix} \quad (4)$$

IIOAB JOURNAL
ISSN: 0976-3104

whereL_xx (x) is the second partial derivative of the image at a point x in the x direction and L_xy (x) is the mixed partial second derivative of the image at a point x in the x and y directions. Then the key point is detected in different direction, location according to the maximum and minimum value with different orientation. The extracted key points are stored in the database by training the features using the compositional neural networks[13]. The network train the feature according to the key point importance and the error has been rectified by updating the Gaussian sigmoid function which is stored in the database. When the new features arriving to the identification process the finger knuckle image is compared with the trained features with the help of the Levenshtein distance which is done as follows,

$$D_{a,b} = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} D_{a,b}(i-1,j)+1 \\ D_{a,b}(i,j-1)+1 \\ D_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases} \quad (5)$$

Where
D_(a,b) is the distance between the user query and template features, 1_((a_i≠b_j)) is the indicator function which is equal to 0 when a_i=b_j
The length of the features like template and user query with respect to the i and j which is used to estimate the distance between the features. Then the computed similarity value is compared with the threshold value 0.3. If the value is greater than the threshold value, then the user related query value is considered as the valid template otherwise leaves as the invalid one. Thus the proposed system extracts the principal curvature based features of the finger knuckle print and those features are trained by Compositional Neural Networks, matching is done with the help of the Levenshtein Distance measure. Then the identification system is implemented by another biometric feature using the Rapid Segment Analysis Feature Transform (RSAFT) method which is explained as follows.

Rapid Segment Analysis Feature Transform (RSAFT) based Identification System

The next identification system is implemented by using the rapid segment analysis feature transform method. Initially ear images are captured and converted into the gray scale image which is done by using the eqn (1). After eliminating the noise from the image, median filter has been applied for removing the noise image which effectively analyze each and every pixel present in the image. The noise has been eliminated by sorting the neighboring pixels[14], if the pixel is corrupted by noise that is replaced by the median value. Once the noise is eliminated by using the rapid segment analysis feature extraction method. This method effectively analyzes the image in terms of using the 16 pixels which helps to determine the key point from the image. From the detected key point, the orientation has been assigned using the magnitude and orientation which is calculated as follows,

$$m(x,y) = \sqrt{\left(L(x+1,y) - L(x-1,y)\right)^2 + \left(L(x,y+1) - L(x,y-1)\right)^2} \quad (6)$$

$$\theta(x,y) = atan2\left(L(x,y+1) - L(x,y-1)\right), \left(L(x+1,y) - L(x-1,y)\right) \quad (7)$$

Where, m(x,y)=magnitude of the key image, θ(x,y)=orientation the key point image
Based on the magnitude value, the direction of the pixel is estimated by using the 36 bin with 360 degree histogram. From the estimated direction [15], the key descriptors are detected by using the candidate image that consists of 16*16 histogram image. The estimated elements are normalized by using the threshold value i.e 0.2 (default threshold value). The key point whose value is within the threshold value as consider as the ear feature. Then the identified ear features are stored in the database which used to match the user query template for further processing. If the new features are entered into the identification system which is compared with the template feature using the Hausdorff distance [16] that is done by as follows,

$$d_H(X,Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y)\} \quad (8)$$

Sup- supremum, inf- infimum.

Where, - Similarity between the training and testing features.
Calculated similarity value is compared with the threshold value 0.2. If the value is greater than the threshold value, the user template considers as the valid template otherwise template is invalid. Thus the proposed system recognizes the ear features by Rapid Segment Analysis Feature Transform and those features are valid by Hausdorff Distance measure. Thus the two biometric features are effectively identifying the user while accessing the particular information. Then the efficiency of the system is analyzed using the following experimental results and discussion.

Infection is a dynamic process involving invasion of body tissues by pathogenic micro-organisms and their toxins. Nosocomial/ hospital/ acquired infections are those which are not present or incubated before admission of patient to the hospital but obtained during the patient's stay in hospital. Lab coats, nurses' uniforms and other hospital garments, materials and articles may play an important part in transmitting pathogenic bacteria in a hospital setting .The hands of healthcare personnel are most commonly implicated in transmitting the pathogens [1]. Various nosocomial pathogens, such as methicillin-sensitive *Staphylococcus aureus* (MSSA), methicillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-resistant *Enterococci* (VRE) and gram negative organisms is well documented [2]. Specifically in the area of dentistry, health care professionals are routinely exposed to potentially pathogenic microorganisms which

IIOAB JOURNAL
ISSN: 0976-3104

are present in the surrounding environment. Most of them originate from the mouths of patients [3]. Contamination may occur from instruments through contamination vectors. These contaminated object infections may be transferred from patient to patient or from patient to professionals [4]. Methicillin resisitant*Staphylococcus aureus* which is the most pathogenic microorganism, comes in contact with health care professionals via direct hand contact with contaminated body fluids, devices, items or environmental surfaces [5].

There are very few studies regarding the wearing and laundering of lab coats in hospitals and medical practice. This study highlights the role of lab coats acting as vector for transmitting health care infections to the patients and the common areas where contamination occurs.

RESULTS AND DISCUSSIONS

In this section analyze the efficiency of the proposed two biometric system in terms of using the false acceptance rate, false rejection rate, equal error rate and accuracy. False Acceptance Rate (FAR) [17] is the process of identifying that rate of unauthorized user acceptance during the identity identification process. The FAR is measured by using the following eqn 9.

$$FAR = \frac{Number\ of\ features\ accepted}{Number\ of\ features\ tested} * 100 \qquad (9)$$

Then the efficiency of the proposed system false acceptance rate is shown in the [Table 1] and the relevant graphical representation is shown in [Fig. 2].

Table 1 : False Acceptance Rate

Methods	Fingerknuckle	ear
SVM	1.65	1.43
NN	1.23	1.14
MLP	0.98	0.86
GGAPC	0.54	0.42
RSAFT	0.32	0.29

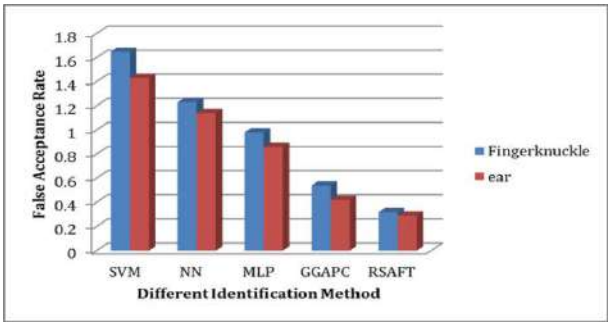


Fig. 2: False acceptance rate.

The [Fig. 2] clearly shows that the two proposed methods consists of minimum false acceptance rate when compared to other identification methods. The minimum false acceptance rate means, the system completely eliminates the unauthorized user while accessing the information from any application. In addition the efficiency of the system is evaluated with the help of the false rejection rate which is shown in the [Table 2] and the relevant graphical representation is depicted in [Fig. 3]. False Rejection Rate [18] is the process of incorrectly rejecting the authorized user during the matching process which was measured in terms of percentage. The FRR is measured by using the following eqn 10.

$$FRR = \frac{Number\ of\ original\ features\ rejected}{Number\ of\ original\ features\ tested} * 100 \qquad (10)$$

Table 2: False rejection rate

Methods	ear	finger knuckle
SVM	0.19	0.23
NN	0.16	0.21
MLP	0.126	0.17
GGAPC	0.072	0.09
RSAFT	0.03	0.02

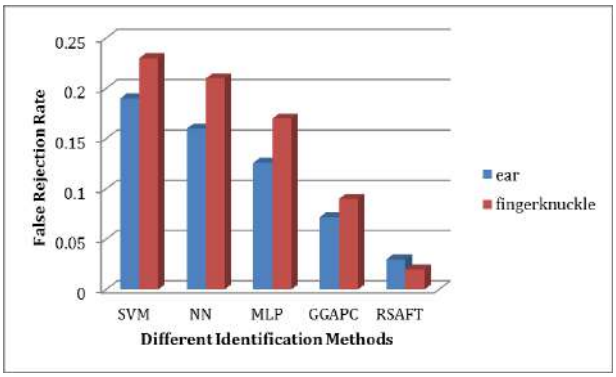


Fig. 3: False rejection rate.

The [Fig. 3] depicted that the false rejection rate of the proposed two methods which is compared with the traditional identification methods. This indicates that the proposed system minimizes the right feature rejection while matching the test and training features. Thus the system consists of equal error rate [19] which is shown in the [Table 3] and the graphical representation is shown in [Fig. 4].

Table 3: Equal error rate

Methods	ear	finger knuckle
SVM	0.046	0.036
NN	0.031	0.024
MLP	0.019	0.012
GGAPC	0.009	0.006
RSAFT	0.004	0.003

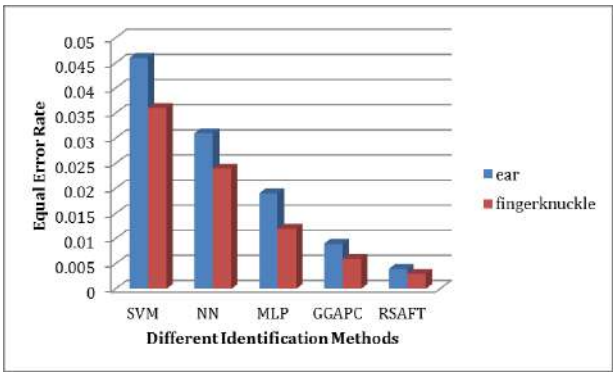


Fig. 4: Equal error rate.

Thus the proposed system successfully recognizes the features which helps to authenticate the user information. The efficiency of the system is analyzed using the following [Table 4] and [Fig. 5].

Table 4: Efficiency

Methods	Ear	finger knuckle
SVM	83	84
NN	88	90
MLP	91	93
GGAPC	97	98
RSAFT	98.43	99.1

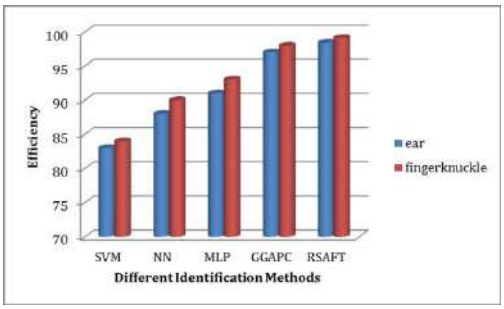


Fig. 5: Efficiency.

From the [Fig. 5], it clearly shows that the proposed system consumes high accuracy [20] for both biometric features, with defined methods. Thus the proposed GGAPC method consumes 97% for ear feature and 98% for finger knuckle biometric feature. In addition the RSAFT method consumes 98.43% for ear feature and 99.1% for finger knuckle biometric feature when compared to the other biometric identification method.

CONCLUSION

Thus the paper compares proposed two biometric based identifications systems such as Graph based Geometric Approach and the Principal Curvature approach and Rapid Segment Analysis Feature Transform (RSAFT) method. These methods utilize the ear and finger knuckle biometric feature for examining the efficiency of the system. This system uses different image processing steps such as color transformation, noise removal process such as non-local median filter median filter and the principle curvature based detector, rapid segment feature extraction process and matching process is done by using the Levenshtein distance and Hausdroff distance measure. Based on the above steps the biometric systems are successfully implemented. The efficiency of the system is evaluated with the help of different performance metrics such as false acceptance rate, false rejection rate, equal error rate, and accuracy metrics. Thus the proposed system ensures the high accuracy when compared to the other identification methods.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
There is no acknowledgement regarding this manuscript.

FINANCIAL DISCLOSURE
There is no financial support for this manuscript.

REFERENCES

[1] A Kumar, Personal identification IITD-BRL-07-2, 2007.
[2] Hongjun Li, Ching Y. Suen, [2015] A novel Non-local means image denoising method based on grey theory, JournalPattern Recognition in ACM.
[3] Jin-Jiang Li, Hui Fan. [2012] Robust Feature Extraction Based on Principal Curvature Direction, Computational Visual Media in Springer.
[4] Kumar. [2014] Importance of Being Unique From Finger Dorsal Patterns: Exploring Minor Finger Knuckle Patterns in Verifying Human Identities, IEEE Transaction on Information Forensics and Security, 9(8).
[5] Mudhafar M. Al-Jarrah. [2012] An Anomaly Detector for Keystroke Dynamics Based on Medians Vector Proximity, Journal of Emerging Trends in Computing and Information Sciences, 3(6).
[6] SayanMaity and Mohamed Abdel-Mottaleb. [2015] 3D Ear Segmentation and Classification Through Indexing, IEEE Transactions on Information Forensics and Security, 10(2).
[7] Vaibhav, Pradeep, [2014] Person Recognition Based on Knuckle Print Biometric Features Computed using Radon Transform, International Journal of Advanced Research in Computer Science and Software Engineering, 4(3).
[8] Mahesh Kumar, Premalatha. [2014] Finger Knuckle-Print Identification Based On Local And Global Feature Extraction Using Sdost, American Journal of Applied Sciences 11 (6): 929-938.
[9] Meraoumia, Chitroub, Bouridane. [2011] Fusion of Finger-Knuckle-Print and Palmprint for an Efficient Multi-Biometric System of Person Recognition, International Conference on Communications in IEEE.
[10] Sukhdev Sing, Chander Kant. [2015] A Multimodal Biometric Identification System Using Finger Knuckle Print and Iris", International Journal of Advanced Research in Computer and Communication Engineering, 4(11).
[11] AsmaaSabetAnwara d, Kareem Kamal A. Ghanyb, HeshamElmahdyc. [2015] Human Ear Recognition Using Geometrical Features Extraction, International Conference on Communication, Management and Information Technology.
[12] S Arastehfar, AA Pouyan, A Jalalian. [2013] An enhanced median filter for removing noise from MR images, Journal of AI and Data Mining, 1(1).
[13] Gajanand Gupta. [2011] Algorithm for Image Processing Using Improved Median Filter and Comparison of Mean, Median and Improved Median Filter, International Journal of Soft Computing and Engineering, 1(5).
[14] Wang Shu-zhong, [2013] An Improved Normalization Method for Ear Feature Extraction, International Journal

of Signal Processing, Image Processing and Pattern Recognition,6(5).
[15] HimanshuMaurya, ShikhaMaurya, "Human Identification by Ear Images using SIFT Algorithm", International Journal of Science and Research, available at, http://www.ijsr.net/archive/v2i5/IJSRON2013978.pdf
[16] Neha Kudu, Dr. Sunil Karamchandani. [2016] Biometric Identification System using Fingerprint and Knuckle as Multimodality Features", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).
[17] Hossain, G Chetty. [2011] Human Identity Verification by Using Physiological and Behavioural Biometric Traits, International Journal of Bioscience, Biochemistry and Bioinformatics, 1(3).
[18] TamijeselvyPerumal, ShilpaSomasundar, [2014] Ear Recognition Using Kernel Based Algorithm, International Journal of Science, Engineering and Technology Research (IJSETR), 3(4).
[19] N.Lavanya Devi. [2015] Human Ear Identification by Haar and Pca, Journal of Research in Computer Science, Engineering and Technology vol 1 Apr 2015.
[20] G.AmarTej ,Prashanth.K.Shah, [2015] Efficient quality analysis and enhancement of MRI image using Filters and Wavelets", International Journal of Advanced Research in Computer and Communication Engineering 4(6).

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

SECURE DATA STORAGE AND SHARING IN CLOUD: VM SCHEDULING

R.G. Babukarthik[1*], J. Satheesh Kumar[2], J. Amudhavel[3]

[1]Part time Research scholar (Category-B) R&D, Bharathiar University, Coimbatore, Tamil Nadu, INDIA
[2]Department of Computer Application, Bharathiar University, Coimbatore, Tamil Nadu, INDIA
[3]Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA

ABSTRACT

Performance of cloud storage is comparatively high compare to other storage devices in terms of consistency, efficiency improvement and execution time. Based on workflow application VM resource type is chosen dynamically in the cloud. A pure isolation hypervisor simply divides a machine into partitions, and permits sharing of resources between the partitions, a load balancing scheme is based on dynamic resource allocation policy for virtual machine cluster, running under para virtualization mode on a cluster of physical machines (PM) in shared storage architecture. In this paper we proposed a mechanism that makes possible for the data owners to enforcement their security policies to ensure data confidentiality and integrity, which enable trusted data sharing through untrusted cloud providers. The computation time for PSO technique is calculated and is compared with various scheduling algorithm such as Round Robin (RR), Random, Heros, Green, Randens and Bestdens.

KEY WORDS
VM Virtual Machine, PM Physical Machine, PSO Particle Swarm Optimization, RR Round Robin

Received: 3 June 2017
Accepted: 20 July 2017
Published: 12 Sept 2017

INTRODUCTION

Workflow applications evolved as an eye-catching paradigm for programming in distributed computing, it is widely used in bioinformatics, scientific computing and physics [1]. Workflow application emerged as a huge data application, due to raising problems of scientific computing systems, leading to vast infrastructures for executing the application within a stipulated time [2]. Thus cloud computing infrastructures pays a special need [3], distributed computing focusing on efficient and cost-effective system operation. Cloud infrastructure is provided on the basis of pay-per-use, facilitating dynamic scaling for large workflow applications. Cloud storage performs well compare to other storage in terms of consistency, efficiency improvement and execution time [4]. Based on workflow application VM resources and its type are chosen dynamically in the cloud. Even though huge amount of resources are available in the cloud, users need to focus on economical cost and its policy [5]. Cost is calculated using time unit model and not on the basis of usage of resources [6, 7]. Generally task of all real world applications differs a lot, exhibiting heterogeneous behaviors (Memory-intensive, data-intensive and computation-intensive) leading to choose an unique VM types [8,9] example E2C of Amazon delivers various instance of VM, such as computing optimized, storage optimized and memory optimized thereby rendering various VM types [10]. Security arise a major concern for various application on cluster [11, 12, 13] heterogeneous distributed computing systems [14, 15], grid computing [16] and cloud computing [17]. International Data Corporation states that cloud security arise as a major concern [18]. This is mainly due to huge amount of user is allowed to execute various untested third party applications, thus applications and users need to be scrutinized [19]. Still many cloud computing environments yet to apply security to tackle the security threats [20]. Hence there is an emerging need to implement security and privacy in cloud data center.

There are very few studies regarding the wearing and laundering of lab coats in hospitals and medical practice. This study highlights the role of lab coats acting as vector for transmitting health care infections to the patients and the common areas where contamination occurs.

RELATED WORKS

Task scheduling is a basic functional unit of cloud architecture from security aspect, it consider various security requirements mainly due to sensitivity of the data such as medical image analysis, image storage and stock photo service. Very high security level is needed in medical image analysis, where as in traditional scheduling consider only makespan and minimizing energy consumption for optimized load balancing [21, 22]. As a result complexity of problems increases gradually; a model encompassing various approaches is stated for security based scheduling [23]. Cloud security services is grouped into three types, application service in the application layer, secure process enabling service (infrastructure layer) and secure physical service (physical layer) based on tools ensuring security (security as a service) [24].

Hypervisor which reduce the overhead of virtualization and provide security: There are two classes of hypervisors that must be considered when examining the technical implications of MLS for hypervisors; they are pure isolation hypervisors and sharing hypervisors. A pure isolation hypervisor simply divides a machine into partitions and permits no sharing of resources between the partitions (other than CPU time and primary memory). Implementing a pure isolation hypervisor is very easy, because the only security policy to be enforced is isolation. IBM's EAL5- evaluated PR/SM system [25] for the z/Series mainframes and it is a good example of a pure isolation hypervisor. The idea of a secure sharing hypervisor originated with Madnick and Donovan [26]. The best examples of such secure sharing hypervisors are KVM/370 [27] and Digital's A1-secure VMM [28, 29, 30]. The most critical feature of a secure sharing hypervisor is a secure shared file store11. The secure shared file store allows a high level partition to have read-only

***Corresponding Author**
Email:
r.g.babukarthik@gmail.com
Tel.: +91-9043108042

IIOAB JOURNAL
ISSN: 0976-3104

access to low-level data, while a low-level partition gets read-write access to the same data. This avoids the clumsy one-way networking approaches only a single copy of the data is required and updates are visible immediately to all partitions [31, 32, 33].

Integrity Using a Virtual Machine Verifier to assure the integrity during the Virtual Machine Mobility: Distributed computing architectures, such as grid and cloud computing, depend on the high integrity execution of each system in the computation [34, 35]. While integrity measurement enables systems to generate proofs of their integrity to remote parties, current integrity measurement approaches are insufficient to prove runtime integrity for systems in these architectures. Integrity measurement approaches that are flexible enough have an incomplete view of runtime integrity, possibly leading to false integrity claims and approaches that provide comprehensive integrity is used only for computing environments that is too restrictive. Proposed architecture is used to build comprehensive runtime integrity proofs for general purpose systems in distributed computing architectures [36, 37]. In this architecture, they strive for classical integrity, using an approximation of the Clark-Wilson integrity model as our target. Key for building such integrity proofs is a carefully crafted host system whose long-term integrity is justified easily using current techniques and few new component called VM verifier, which comprehensively enforces our integrity target on VMs. Building a prototype based on the Xen virtual machine system for SELinux VMs and to find the distributed compilation is implemented, thereby providing accurate proofs of our integrity target with less than 4% overhead [30].

PROPOSED WORK

Towards secure data storage and sharing in the cloud: End-users of the cloud store their data in the provider's infrastructure; a critical concern is the security and privacy of these data. End-users want to know where their data is stored and who has control of the data in addition to the owners. They also want to be protected against illegal access to the data by the cloud provider, or other third parties. Secure access and storage of data in the cloud is addressed through the following tasks.

Data leakage prevention and privacy with 3rd party service providers: Shifting data storage to off-premises providers has two consequences: First, data owners have only limited control over the data stored in the cloud. Second, cloud providers have excessive privileges, giving them extensive control over the cloud user's data. Combining, this leads to a low level of trust between the end-user and the cloud provider with respect to keeping and sharing business critical data in the cloud. Mechanisms that make it possible for the data owners to enforcement their security policies to ensure data confidentiality and integrity, mechanisms that enable trusted data sharing through untrusted cloud providers.

Information source authentication: Algorithms that guarantee the authenticity of data stored in the cloud. This provide authentication and trust in the acquired information to avoid situations where the user's data may be altered without the owner's consent.

Methods to enable free inter cloud data movement: End-users buy services offered by the cloud providers without knowing where the cloud resources are located. The location might be beyond the borders of a legislative entity and can cause problems when disputes happen, which might be beyond the control of cloud provider. Furthermore, entrusting significant amounts of data to a cloud provider creates a risk of data lock-in. Technical solutions that ease the implementation of free inter- cloud data movement and a policy specification to standardize that process between providers.

Policies for data retention: Data retention is defined as storing recorded data for a period of time that is longer than the time necessary to perform the tasks; this remains the reason for recording the data. E.g. Amazon stores the list of all previous purchases for each individual even though it is not necessary for practical or legal reasons (order completion, accounting etc.). Currently, data retention is usually regulated in the terms of a service agreement between the provider and the user. Due to the complexity and frequent changes of such agreements they are usually not read by the user before they are accepted, developed policies for data retention and technical tools for enforcing these policies.

End results and significance: The security solutions increase the trust between cloud customers and cloud providers, thereby increasing the security of the cloud services and infrastructure. Moreover, sensitive data can be securely stored, shared and processed in the cloud. This allows businesses to reap the full benefits of the cloud, and a business critical decision is made with all relevant data available.

PSO VM scheduling algorithm is used for optimal scheduling not only focusing on makespan and energy efficient scheduling but also on the security and privacy principles. The number of servers is given as input keeping the fixed number of user and specific topology. The first step starts with the initialization of server, user, topology and data center components. The second step is setting of parameters such as power models, core switch and aggregate switch. The execution of task is taken place using PSO scheduling algorithm and with other scheduling algorithm the performance of proposed scheduling algorithm is compared. The computation time and energy is calculated.

IIOAB JOURNAL
ISSN: 0976-3104

PSO_VMscheduling_Algorithm
Input: No. of servers, topology, user.
Output: Energy, computation time, memory.
Step 1: Initialization
Server, topology, user, data center components.
Step 2: Parameters Setting
Non-linear power model, core switch, aggregation switch, task.
Step 3: Execution
Scheduler
PSO, Round Robin, Random, Heros,
RandDENS, BestDENS
Topology
Three-tier debug
Energy model
eDVFS, DNS.
Step 4: Report
Data center load
Individual server load
Individual VM load
Load of individual links
Step 5: Display
Total energy consumed
Energy of servers

EXPERIMENTAL EVALUATION AND ANALYSIS

The experiment is carried out using the green cloud simulation tool, the input is given as number of server, for a given fixed number of user and topology of the network, the computation time for PSO technique is calculated and is compared with various scheduling algorithm such as Round Robin (RR), Random, Heros, Green, Randens and Bestdens. It is clear that computation time for green scheduling is minimum compared to the all other scheduling algorithm at minimum number of server (tasks), whereas as the number of server increases PSO really out performed well. The [Table 1] shows the computation time for various scheduling algorithm. [Fig. 1] show the comparison of the computation time for various scheduling algorithm and thus PSO really out performed well, underlying below the all scheduling algorithm with minimum computation time.

Table 1: Computation time comparison of various scheduling algorithm

Sl. No.	Server	RR	Random	Heros	Green	Randens	Bestdens	PSO
1	30	0.5	0.41	1	0.36	0.81	0.89	0.42
2	60	1	1	2	1	1	2	1
3	90	2	2	4	2	2	3	2
4	120	3	3	7	4	3	5	3
5	150	4	4	10	5	4	8	4
6	180	4.5	5	13	6	5	10	4.4
7	210	5	5.7	17	8	6	13	5
8	240	6	6	22	9	8	16	6
9	270	7	7	28	11	9	20	7
10	300	8	9	35	13	10	23	8

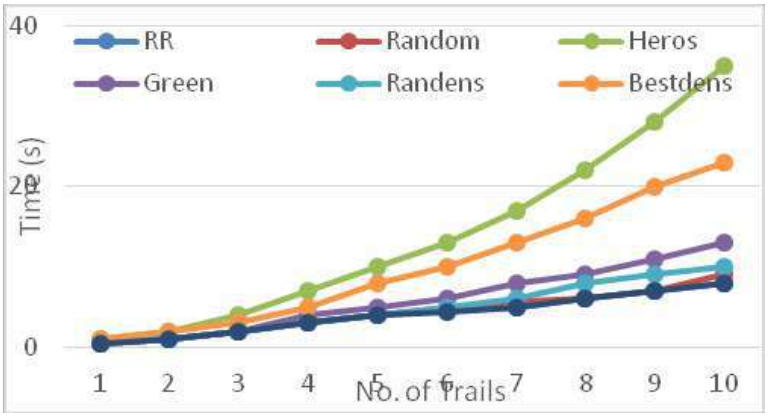


Fig. 1: comparison of scheduling algorithm.

CONCLUSION

Cloud infrastructure is provided on the basis of pay-per-use, facilitating dynamic scaling for large workflow applications. Secure access and storage of data in the cloud is addressed by Data leakage prevention and privacy with 3rd party service providers, Information source authentication, various methods to enable free inter cloud data movement, Policies for data retention and End results significance. The computation time for PSO technique is calculated and is compared with various scheduling algorithm, green scheduling has minimum computation time compared to the all other scheduling algorithm with minimum number of server (tasks), whereas as the number of server increases PSO really out performed well.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Juve G, Chervenak A, Deelman E, Bharathi S, Mehta G, Vahi K.[2013] Characterizing and profiling scientific workflows, Future Generation Computer System 29(3): 682 - 692.

[2] Kashlev A, Lu SY.[2014] A system architecture for running big data workflows in the cloud, 2014 IEEE International Conference on Services Computing, SCC pp. 51-58.

[3] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I,[2009] Cloud computing and emerging it platforms: vision, hype, and reality for delivering computing as the 5th utility ,Future Generation Computer System 25(6): 599 -616.

[4] Chang V, Wills G.[2015] A model to compare cloud and non-cloud storage of Big Data, Future Generation Computer Systems.

[5] Calheiros RN, Buyya R.[2014] Meeting deadlines of scientific workflows in public clouds with tasks replication, IEEE Transactions on Parallel and Distributed Systems 25(7) :1787-1796.

[6] Rodriguez MA, Buyya R.[2014] Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds, IEEE Transactions on Cloud Computing 2(2):222-235.

[7] Fard HM, Fahringer T, Prodan R.[2013] Budget - constrained resource provisioning for scientific applications in clouds, 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, Cloud Com pp. 315-322.

[8] Thomas G, Williams AB.[2009], Sequential auctions for heterogeneous task allocation in multiagent routing domains, IEEE International Conference on Systems, Man and Cybernetics, SMC pp. 1995-200.

[9] Iturriaga S, Nesmachnow S, Luna F, Alba E.[2015] A parallel local search in CPU/GPU for scheduling independent tasks on large heterogeneous computing systems, Journal of Supercomputing 71 (2):648 -672.

[10] Amazon EC2, http://aws.amazon.com/ec2/, 2015.

[11] Amin A, Ammar R, El Dessouly A.[2004] Scheduling real time parallel structures on cluster computing with possible processor failures, Ninth International Symposium on Computers and Communications, ISCC pp. 62 -67.

[12] Apvrille A, Pourzandi M.[2004] XML distributed security policy for clusters, Computers & Security 23(8):649-658.

[13] Xie T, Qin X.[2006] Scheduling security -critical real -time applications on clusters, IEEE Transactions on Computers 55(7) :864 -879.

[14] Song S, Hwang K, Kwok YK. [2006] Risk-resilient heuristics and genetic algorithms for security -assured grid job scheduling, IEEE Transactions on Computers 55(6):703 –719.

[15] Xie T, Qin X.[2007] Performance evaluation of a new scheduling algorithm for distributed systems with security heterogeneity, Journal of Parallel and Distributed Computing 67(10) :1067 -1081.

[16] Tang X, Li K, Zeng Z, Veeravalli B. [2011] A novel security -driven scheduling algorithm for precedence constrained tasks in heterogeneous distributed systems, IEEE Transactions on Computers 60(7):1017-1029.

[17] Zeng LF, Veeravalli B, Li XR. [2015] SABA: a security - aware and budget-aware workflow scheduling strategy in clouds, Journal of Parallel and Distributed Computing 75: 141 -151.

[18] Gens F., IT cloud services user survey, pt.2: Top benefits & challenges (October 2008). URL http://blogs.idc.com/ie/?p=210
[19] Yurcik W, Meng X, Koenig G, Greenseid J. [2004] Cluster security as a unique problem with emergent properties, Fifth LCI International Conference on Linux Clusters: The HPC Revolution 2004, May 2004.
[20] Behl A, Behl K.[2012] An analysis of cloud computing security issues, 2012 World Congress on Information and Communication Technologies, WICT 109 -114.
[21] Magoulès F, Pan J, Teng F. [2012] Cloud Computing: Data-intensive Computing and Scheduling, CRC press,
[22] Kolodzie j, Xhafa F.[2011] Meeting security and user behavior requirements in grid scheduling, Simul. Modell.Pract. Theory19(1):213–226,doi: 10.1016/j.simpat.2010.06.007.
[23] Khan AN, Mat Kiah ML, Khan SU, Madani SA. towards secure mobile cloud computing: a survey, Future Gener. Comput. Syst.29 (5) (2013) 1278–1299, doi: 10.1016/j.future.2012.08.003.
[24] A Furfaro, A Garro, A Tundis, towards security as a service (secaas): On the modelingof security services for cloud computing, in: 2014 International Carnahan Conference on Security Technology (ICCST), 2014, pp. 1–6, doi: 10.1109/CCST.2014.6986995.
[25] Certification Report for Processor Resource/ System Manager (PR/SM) for the IBM eServer zSeries 900, BSI-DSZ-CC-0179-2003, 27 February 2003, Bundesamt für Sicherheit in der Information stechnik: Bonn, Germany. URL:http://www.commoncriteriaportal.org/public/files/epfiles/0179a.pdf
[26] Madnick SE, Donovan JJ. Application and Analysis of the Virtual Machine Approach to Information System Security. In Proceedings of the ACM SIGARCH-SIGOPS Workshop on Virtual Computer Systems. 26-27 March 1973, Cambridge, MA: Association for Computing Machinery. p. 210-224. URL: http:// portal.acm.org/ citation.cfm?id=803961.
[27] M. Schaefer,, B. Gold, R. Linde, and J. Scheid. Program Confinement in KVM/370. In Proceedings of the 1977 ACM Annual Conference. 16-19 October 1977, Seattle, WA: p. 404-410.
[28] Karger PA, PA ME Zurko, Bonin DW, AH Mason, Kahn CE, A Retrospective on the VAX VMM Security Kernel. IEEE Transactions on Software Engineering, November 1991. 17(11): 1147-1165.
[29] Joshua Schiffman, Thomas Moyer, Christopher Shal [2009]Justifying Integrity Using a Virtual Machine Verifier" IEEE Annual Computer Security Applications Conference
[30] Anusha B, Noah, Sivaranjani C, Priyanka S, [2015.]Predictive analysis of movie reviews using hybrid approach", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 1(1): 1-7
[31] Dharshini G, Subhasri V, Sujitha G, Ganesan M, [2016] "Secure Information Retrival for Decentralised Disruption Tolerant Military Networks using CP-ABE", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(1): 1-6
[32] Govindharaj I, Karthiga S, Manishalakshmi R.[2016] R Mary Silvia Theodore, "Home Power Analyzer with Smart Power Monitoring using IoT", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(1): 7-13
[33] Ahilandeswari T, Nandhini S, Sivasankari P, Rajalakshmy M.[2016] Intensifying the Generic Middleware for Smart Environment, International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(2): 1-5,
[34] Gayathri R, Indumathi K, Githanjali P, Roobini V, [2016] Securing Multimedia using Data Lineage in Malicious Environment: A Survey, International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(.2): 22-29.
[35] Shanmugam M, Dhivya S, Lavanya B, Keerthana V.[2017] Free Voice Calling in Wi-Fi Network using Android, International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 3(1):1-8
[36] Rajadurai R, Amelia, Aubrey, A.Anusha, Danapriya P. Geethashnee D.[2017] Efficient Data Leakage Prevention Strategy using Key Distribution", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 3(1): 9-16
[37] Sasidevi V, Hannah D Sathiyan, Rajadurai R.[2017] Classification Algorithm: A Survey", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 3(1): 7-21

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

TO SOLVE ECONOMIC DISPATCH PROBLEM USING COOPERATIVE PARTICLE SWARM OPTIMIZATION ALGORITHM

T. Asvany[1*], J. Amudhavel[2], Sujatha Pothula[3]
[1,3]*Department of Department of Computer Science, Pondicherry University, Pondicherry, INDIA*
[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT
An efficient Cooperative Particle Swarm Optimization (CPSO) algorithm is used to solve the Economic Dispatch (ED) problem in power systems. Generally, PSO technique helps to explore the search space very efficient and effectively. The presented CPSO algorithm is very effective to solving the constraints and the objective function of ED problems in power systems. The results of proposed algorithm are compared with existing algorithms like Genetic algorithm, Tabu search technique and evolutionary programming approaches. The experimental result indicates the proposed CPSO technique was really capable to achieve higher quality solutions in ED problems.

KEY WORDS
Economic Dispatch (ED), Cooperative Particle Swarm Optimization (CPSO), Power system, Evolutionary Programming, Genetic Algorithm.

Received: 5 June 2017
Accepted: 25 July 2017
Published: 14 Sept 2017

***Corresponding Author**
Email:
asvanytandabani@gmail.com

INTRODUCTION
Economic dispatch (ED) problem is one of the popular and most important and basic essential issues in power system operation. In nature optimization problems are very interesting and their characteristics (both complex and nonlinear) have satisfying constraints (equality and inequality). In recent world, there are so many mathematical approaches used, such as Tabu search method, genetic algorithm, Particle swarm optimization (PSO), Ant colony optimization (ACO), and simulated annealing. The swarm intelligence based PSO algorithm is very powerful and produce optimal solutions to global optimal solutions in power system problems in optimization [1].

The Economic Dispatch (ED) problem is previously discussed in various different techniques includes conventional methods, lambda iteration method, gradient method, base point technique and participation factors method [2, 3, 4, 5]. The main aim of optimization concept and their objective is used to minimize the overall generation cost of units, and also satisfying constraints with help of mathematical modeling and used optimization techniques. The ED problem main objective is to find out an optimal combination of power productivity to meet the customer requirement (demand) at minimum cost, at the same time satisfying the constraints also. For ease, the ED problem each unit cost function has been approximately stand for single quadratic function and solved by use of mathematical techniques [6]. In general mathematical methods need some derivative information of the cost function. The generating units of input and output characteristics are non-convex due to their prohibited operating zones, multi-fuel effects and valve point loadings etc.Dynamic programming (DP) technique is used to solve the ED problems with valve-point modeling have been presented by [2, 3]. This dynamic programming method is difficult to solve the dimensionality to become extremely large, and thus it requires huge computational efforts.

Particle Swarm Optimization (PSO) algorithm is one of the most powerful algorithms in modern heuristic world, this algorithm is suitable to solve simple and also large scale non-convex optimization problems. PSO algorithm is developed by Eberhart and kennedy based on the idea of analogy of swarm of bird and fish schooling [7]. The PSO technique imitate swarm individuals behavior and to maximize the species survivals. In PSO technique, each individual makes his own decision using their own and neighbors experience together [8]. This algorithm is basically a population based search algorithm, and searches a space by parallel with the help of group of particles moving across in a multidimensional search space. Each individual particle are stochastically moving toward the position of current best velocity of each individual particle, and their own previous best performance value and also the neighbors previous best performance value [9].

The important and most powerful advantage of PSO algorithm is very simple in concept, implementation is easy, minimum number of parameters is used, robustness in nature, computational efficiency is very high compared to other mathematical models and other heuristic based optimization methods. The searching capabilities of PSO algorithm is good and it may get trapped in local minimum while managing heavily constrained problems due to the limited global/local searching capabilities [10, 11].

In early years, Genetic Algorithm (GA) or Simulated annealing (SA) techniques are comes under global optimization technique. A probabilistic heuristic algorithm is used to solve successfully in a power optimization problems such as feeder configuration & capacitor placement in a distributed system [12-15]. Compared to simulated annealing technique the GA method is usually fast, for the reason that GA generally has a parallel search technique. Due to their natural genetic operations, global optimization has high potential and GA techniques have great effect in solving ED problems. Using GA technique lot of changes done in the ED problems like ramp rate limits, network losses and valve point zone are also considered.

IIOAB JOURNAL
ISSN: 0976-3104

Authors walters and Sheble is used a GA technique to solve the ED problem for valve-point discontinuities that employed unit output as encoded in parameter of chromosome [16]. Chen and Chang also presented a same general GA method to solve ED problems in a system incremented cost as encoded parameter can include the ramp rate limits, network losses and valve-point zone [17]. Fung et al. presented a GA method incorporating with Tabu search (TS) and Simulated Annealing (SA) methods that employed the generators output as encoded parameter [18]. Yalcinoz et al. used an efficient GA technique in the real coded representation scheme, generally the GA technique have crossover (arithmetic), mutation and elitism in GA is used to solve more efficiently for ED problems in a high quality solution obtain with less computation time period [19].

This paper presented a PSO based approach to solve the Economic Dispatch (ED) problems with heavy constraints. To overcome existing issues, made some changes and combining techniques in a sequence. Additionally, the crossover technique is proposed to improve the solution quality without scarifying the efficiency of computational time. The main aim of this paper is to solve ED problems in an existing CPSO technique [21].

ECONOMIC DISPATCH PROBLEM

In modern world a good business practice is one which minimizes the production cost at the same time without sacrificing the quality. In power system generations has different number of power plants and each power plant system has different number of generating units. At any time, the total load system will met in the generating units in different power plants system. To determine the Economic Dispatch problem control the power output of each power plant system and generating unit of each power system unit within power plant system, which is used to minimize the total cost of fuel required to serve the system load.
The main objective of Economic Dispatch (ED) problem is to minimize the total generation cost and also satisfies the equality and inequality constraints, to meet the power system load demand. The ED problem is to find the each power system have the real power generation such as the objective function (i.e., overall production cost function) as defined by the equation 1.

$$C_t = \sum_{i \in I}^{n} F_i(P_i) \quad (1)$$

The transmission losses is a major factor to transmitted power in long distances and affect the optimum best dispatch of generation. One common thing they include to express an effect of transmission loss in an overall transmission loss of the power output generators with the help of quadratic function. The simplest representation of the quadratic equation 2 is

$$F_i(P_i) = a_i + b_i P_i + c_i P_i^2 \quad (2)$$

Where

C_t - Overall generation cost;

F_i - Generator ith Cost function;

a_i , b_i, c_i - three cost coefficients of generator i;

P_i - Generator ith Electrical output;

i - Set value for all generators.

The overall generation cost is minimizing, the overall generation cost should be equal to total number of system demand and transmission network loss. In this work network loss is not considered, cost function is detailed description is listed. It is minimum subject to the constraints are represented equation 3

$$\sum_{i \in I}^{n} P_i = P_D \quad (3)$$

Where P_D is the overall system demand. Subject to constraints, number of generations is equal to overall demand plus their respective losses, i.e, equation 4 is shown below

$$\sum_{i \in I}^{n} P_i = P_D + P_L \quad (4)$$

Where the P_D represents the overall system is load and P_L is the total transmission network loss. The power output of each generator unit should be within its minimum and maximum limits. Here they satisfying the inequality constraints are expressed in the way as follows equation 5

$$P_i,min \leq P_i \leq P_i,max \quad (5)$$

Where P_i,min and P_i,max both represented as minimum and maximum output of generator i respectively.

OVERVIEW OF PARTICLE SWARM OPTIMIZATION

In the year 1995, the PSO algorithm was first introduced by Kennedy and Eberhart [7], this concept is admirable by social behavior of swarm individuals (each particle) such as bird flocking and fish schooling. The PSO provides a population based search technique, each individual represented as particles and each particle move in a multidimensional search space. In a given search space each individual particle adjust its own position according to its own experience of particle, and their neighboring experience of other particle to their best position.

IIOAB JOURNAL
ISSN: 0976-3104

Let us assume the n- dimensional search space, the position (particle co-ordinates) and their velocity of individual (particle j) are represented as $V_j = (v_{j1}, \ldots . v_{jn})$ and $X_j = (x_{j1}, \ldots . x_{jn})$ vectors in the PSO algorithm. the j[th] individual particle previous best position is $Pbest_j = (x_{j1}^{pbest}, \ldots . x_{jn}^{pbest})$ stored in $Pbest_j$. The best particle value among all the particle in the group of individual and its global best value is updated in $Gbest_j = (x_{j1}^{gbest}, \ldots . x_{jn}^{gbest})$. The each particle position and velocity is modified, and calculated by best velocity and their distance from $Pbest_{jn}$ to $Gbest_{jn}$as described in the following formulas 6, 7 and 8:

$$V_{jn}^{i+1} = w.V_{jn}^{i} + c_1 * rand() * (pbest_{jn} - x_{jn}^{i}) + c_2 * rand() * (gbest_n - x_{jn}^{i}) \quad (6)$$

$$x_{jn}^{i+1} = x_{jn}^{i} + V_{jn}^{i+1}, j = 1, 2, \ldots, m \quad (7)$$

$$n = 1, 2, \ldots l \quad (8)$$

Where

m - Total number of individual in a group

I - Total number of members in a individual (Particle)

i - Iterations of pointer (generations)

w - Inertia weight component

c_1, c_2 - Constant (Acceleration constant)

$rand()$ - Random uniform value, Range between [0, 1]

V_j^i - Individual j[th] velocity at iteration i

x_j^i - Individual j[th] current position at iteration i.

The presented above equations demonstrated the searching mechanism of PSO algorithm using position and their velocity of each individual.

COOPERATIVE PSO FOR ED PROBLEMS

In this sector, a Cooperative PSO algorithm is presented to derive Economic Dispatch (ED) problems in power system. Here both the constraints of equality and inequality are also satisfied with the help of this technique to solve ED problems. Especially, the multidimensional search space is used to increase the convergence speed and the each individual search point is devised. The CPSO algorithm can be presented as follows:

Procedure of CPSO

Step 1: Random initialization

Initialize random position of all individuals $X^{sl} = (x_1^{sl}, x_2^{sl}, \ldots . x_m^{sl})$ of slave swarm of size $m^{sl} = m$.

Initialize randomly the position of all individuals $X^{ma} = (x_1^{ma}, x_2^{ma}, \ldots . x_m^{ma})$ of master swarm of size $m^{ma} = m$.

Initialize randomly the velocity of all individuals $V^{sl} = (v_1^{sl}, v_2^{sl}, \ldots . v_m^{sl})$ of slave swarm.

Initialize randomly the velocity of all individuals $V^{ma} = (v_1^{ma}, v_2^{ma}, \ldots . v_m^{ma})$ of master swarm.

Evaluate the fitness value of X^{sl} and X^{ma}.

Set X^{ma} to be $pbest^{ma} = (pbest_1^{ma}, pbest_2^{ma}, \ldots . pbest_m^{ma})$for each individual of the master swarm.

Set the particle with best fitness to be $gbest$.

Set the particle with best fitness of slave swarm to be $gbest^{sl}$.

Set generation i = 0.

Step 2: Reproduction and loop updating

Step 2.1: Slave swarm update

for i = 1 : m

for j = 1: N

Update the velocity v_i^{sl} of individual x_i^{sl} using the below equations.

$$v_{ij}^{sl}(k+1) \leftarrow c_1^{sl} r_1 (1 - r_2)(x_{tj}^{sl}(k) - x_{ij}^{sl}(k) + c_2^{sl}(1 - r_1) r_2 (gbest_j(k) - x_{ij}^{sl}(k))$$

$$v_{ij}^{sl} \leftarrow \min (v_{ij\,max}^{sl}, \max (v_{ij\,min}^{sl}, v_{ij}^{sl}))$$

Update the position of individual x_i^{sl} using the below equation

$$x_{ij}^{sl}(k+1) \leftarrow x_{ij}^{sl}(k+1) + v_{ij}^{sl}(k)$$

If $x_{ij}^{sl} > x_{ij\,max}^{sl}$,set $x_{ij}^{sl} \leftarrow x_{ij\,max}^{sl}, v_{ij}^{sl} \leftarrow v_{ij}^{sl}$;

If $x_{ij}^{sl} < x_{ij\,min}^{sl}$, set $x_{ij}^{sl} \leftarrow x_{ij\,min}^{sl}, v_{ij}^{sl} \leftarrow v_{ij}^{sl}$.

End for
End for
Update $gbest^{sl}$.
Step 2.2: Master swarm Update
for i = 1 : m
for j = 1: N
Update the velocity v_i^{ma} of individual x_i^{ma} using the following equations.

$$v_{ij}^{ma}(k+1) \leftarrow w^{ma}.v_{ij}^{ma}(k) + c_1^{ma} r_1 (1-r_2)(1-r_3)\left(pbest_{ij}^{ma}(k) - x_{ij}^{ma}(k)\right) + c_2^{ma} r_2 (1-r_1)(1 - r_3\left(gbest_j^{ma}(k) - x_{ij}^{ma}(k)\right) + c_3^{ma} r_3 (1-r_1)(1-r_2)(gbest_j(k) - x_{ij}^{ma}(k)))$$

$$v_{ij}^{ma} \leftarrow \min\left(v_{ij\,max}^{ma}, \max\left(v_{ij\,min}^{ma}, v_{ij}^{ma}\right)\right)$$

Update the position of individual x_i^{ma} using the following equation

$$x_{ij}^{ma}(k+1) \leftarrow x_{ij}^{ma}(k) + v_{ij}^{ma}(k+1)$$

If $x_{ij}^{ma} > x_{ij\,max}^{ma}$,set $x_{ij}^{ma} \leftarrow x_{ij\,max}^{ma}, v_{ij}^{ma} \leftarrow v_{ij}^{ma}$;
If $x_{ij}^{ma} < x_{ij\,min}^{ma}$, set $x_{ij}^{ma} \leftarrow x_{ij\,min}^{ma}, v_{ij}^{ma} \leftarrow v_{ij}^{ma}$.
End for
End for
Update $pbest_i^{ma}$
End for
Update $gbest$.
Set $k = k + 1$.
Step 3: if not met termination condition, go to step , otherwise end CPSO

Generally, simultaneously handle two tasks are very difficult for swarm, and these improved methods to make a fast convergence speed and distribution of diversity among the individuals. In this mechanism the new CPSO is introduced and two swarm techniques are used, one is slave swarm and another one is master swarm (the two swarms are used instead of previous swarms) [21]. The slave swarm only first focal point on exploiting the search interval to ignore trapping into local optima and to retain the diversity of individuals.

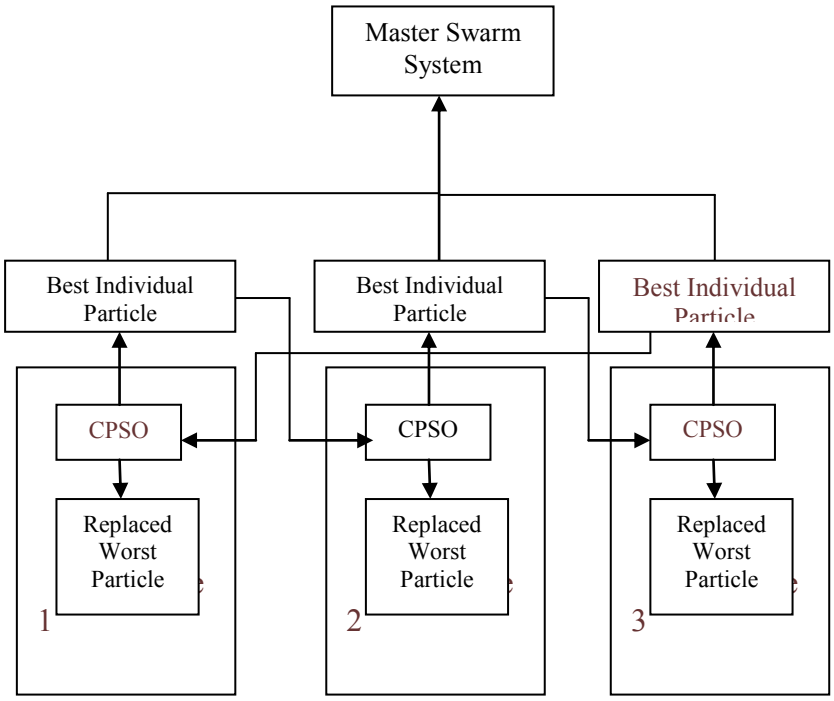


Fig.1: The working procedure of CPSO.

The CPSO algorithm was mainly utilized to establish the best optimal generation power of each unit that was present at the specific period, and thus minimizing the total cost generation. Generally the PSO techniques have three components used; they are namely inertial weight, cognitive parameters, and social parameters. The inertial weight element reproduces the inertial behavior of birds flying previous direction.

The two parameter used, cognitive used to maintain the birds memory in best previous position and social is used to maintain best position among the individuals (the parallel interaction between the swarms). In the dynamic search space where the best optimal solution is required, each individual in the swarm is moved from the optimal point by adding together a velocity with its position. Based on the mathematical model the velocity update equation 9 is

$$v_{ij}(k+1) = c_1 r_1 (pbest_{ij}(k) - x_{ij}(k)) + c_2 r_2 (gbest_j(k) - x_{ij}(k)) \quad (9)$$

For high level convergence speed, the presented method is used to update the velocity of the individuals[32-35], the equation 10 reprsented below.

$$v_{ij}(k+1) = c_1 r_1 \left(x_{ij}(k) - x_{ij}(k)\right) + c_2 r_2 (gbest_j(k) - x_{ij}(k)) \quad (10)$$

The [Fig. 1] represents the overall working flow of CPSO algorithm, here the swarm master maintains the slave swarms best particle value and again sends the best value to slave swarm 1 to slave swarm2 like that. The strategy represented in this to balance exploitation and exploration of presented approach.

RESULTS AND DISCUSSION

An efficiency to assess the proposed CPSO algorithm is developed and used in ED problems; the main objective function is also compared with their specific constraints. Here three benchmark test systems are taken, they are 6-units, 20-units, and 38-units are considered. The setting of parameter constraints is same for all the test systems and maximum number of iterations is excluded.
The power systems have the reasonable B loss coefficients matrix value and that was working to represent transmission loss and their constraints. The code was implemented in Matlab language and executed in the personal computer.

Test System 1: Six unit system

The Test case system contains six benchmark units. The overall system load is 1263 MW. In the six unit power generator outputs, such as listed P1, P2, P3, P4, P5 and P6 are randomly generated. The data input are taken from six generating unit system [29]. The proposed methods of the CPSO and their other comparing algorithms best solutions are described in [Table 1]. The [Table 1] results are compared with SA [28], TS [28], PSO [28], GA [30], and NPSO – LRS [31]. The [Table 1] listed the statistical values

Table 1: Best solution of six unit system

Method	Min (S/hr)	Mean (S/hr)	Max (S/hr)	TFE	Time (sec.)	Standard Deviation
SA	15461.1	15488.98	15545.5	NA	50.36	28.367
TS	15454.89	15472.56	15498.05	NA	20.55	13.719
PSO	15450.14	15465.83	15491.71	100,000	6.82	10.15
GA	15459	15469	15524	20,000	41.58	0.057
NPSO-LRS	15450	15454	15492	20,000	14.89	0.002
CPSO	15443.2	15458	15490	20,000	14.12	0.0018

The [Table 1] observed values for Minimum, Mean for optimal solution to generate the six unit system. The standard deviation results of the proposed system are 0.0018. Compared to other existing taken algorithms the results are shown in the above table.

Test System 2: Twenty unit system

The data input for twenty generating units system is taken from [24]. The [Table 2] listed a detailed test system unit of 20 generators; the load demand value is 2500 MW. The results are compared with BBO [25], LI [26], HM [26], PSO [27], and IPSO [27].

Table 2: Best solution of twenty unit system

Unit Number	BBO	LI	HM	PSO	IPSO	CPSO
P1	513.09	512.78	512.78	270.2587	483.1617	221.6283
P2	173.35	169.10	169.10	184.0766	127.9918	118.5223
P3	126.92	126.89	126.89	50	58.93857	50

IIOAB JOURNAL
ISSN: 0976-3104

P4	103.33	102.87	102.87	70.93799	63.08068	91.75936
P5	113.77	113.64	113.68	61.38516	101.6452	64.39798
P6	73.07	73.57	73.57	20	53.40166	20
P7	114.98	115.29	115.29	118.0858	120.7673	118.1018
P8	116.42	116.40	116.40	50	50	50
P9	100.69	100.41	100.41	147.2164	70.47716	155.5842
P10	100.00	106.03	106.03	80.26374	40.14163	62.84119
P11	148.98	150.24	150.24	241.0485	238.8903	261.1216
P12	294.02	292.76	292.76	400.8137	434.4287	422.8988
P13	119.58	119.12	119.12	96.81035	113.1663	97.7352
P14	30.55	30.83	30.83	93.79342	80.23303	88.89034
P15	116.45	115.81	115.81	62.45174	104.5378	77.67747
P16	36.23	36.25	36.25	38.51789	44.24926	33.31954
P17	66.86	66.86	66.86	33.0497	57.2506	30
P18	88.55	87.97	87.97	30	56.86099	41.80909
P19	100.98	100.80	100.80	87.68055	54.10037	85.5325
P20	54.27	54.31	54.31	34.15818	30	30
Total Generation (MW)	2592.10	2591.97	2591.97	2170.548	2383.323	2121.82
Total Transmission Loss (MW)	92.10	91.97	91.97	329.4517	116.6769	378.18
Total Generation cost (S/h)	62456.78	62456.64	62456.63	60213	60221.96	60199.45

Test System 3: Thirty Eight unit system

The CPSO algorithm is applied to Economic Dispatch problems, here 38 generators unit system with cost and transmission loss. The [Table 3] listed the detailed test system values of 38 generator system unit. The data input values are taken from [21] and load demand is 6000 MW. The results are compared with DE/BBO [22], BBO [22], PSO-TVAC [23], NEW-PSO [23] and EP-EPSO [21].

Table 3: Best solution of thirty eight unit system

Unit Number	DE/BBO	BBO	PSO-TVAC	NEW-PSO	EP-EPSO	CPSO
P1	426.60606	422.230586	443.659	550	318.0777	497.4661
P2	426.606054	422.117933	342.956	512.263	475.117	324.1907
P3	429.663164	435.779411	433.117	485.733	399.1265	326.988
P4	429.663181	445.48195	500	391.083	500	500
P5	429.663193	428.475752	410.539	443.846	500	327.0108
P6	429.663164	428.649254	492.864	358.398	500	326.5769
P7	429.663185	428.119288	409.483	415.729	500	327.4176
P8	429.663168	429.900663	446.079	320.816	500	327.0777
P9	114	115.904947	119.566	115.347	114	114
P10	114	114.115368	137.274	204.422	132.7826	114
P11	119.768	115.418662	138.933	114	114	114
P12	127.0728	127.511404	155.401	249.197	114	114
P13	110	110.000948	121.719	118.886	110	110
P14	90	90.0217671	90.924	102.802	90	90
P15	82	82	97.941	89.039	82	82
P16	120	120.038496	128.106	120	120	120
P17	159.598	160.303835	189.108	156.562	141.9435	147.1996
P18	65	65.0001141	65	84.265	65	65.00002
P19	65	65.000137	65	65.041	65	65
P20	272	271.999591	267.422	151.104	120	272
P21	272	271.872268	221.383	226.344	272	272
P22	260	259.732054	130.804	209.298	260	260
P23	130.648618	125.993076	124.269	85.719	80	96.77796
P24	10	10.4134771	11.535	10	10	10
P25	113.305034	109.417723	77.103	60	92.9577	85.36166
P26	88.0669159	89.3772664	55.018	90.489	55	72.1951
P27	37.5051018	36.4110655	75	39.67	35	35
P28	20	20.009888	21.682	20	20	21.19425
P29	20	20.0089554	29.829	20.985	20	20
P30	20	20	20.326	22.81	20	20
P31	20	20	20	20	20	20
P32	20	20.0033959	21.84	20.416	20	20
P33	25	25.0066586	25.62	25	25	25
P34	18	18.0222107	24.261	21.319	18	18

P35	8	8.0000426	9.667	9.122	8	8
P36	25	25.006066	25	25.184	25	25
P37	21.782	22.0005641	31.642	20	38	20
P38	21.0621792	20.6076309	29.935	25.104	20	20
Total Cost	9,417,235.79	9,417,633.64	9,500,448.31	9,596,448.31	9,387,925.50	9013940

CONCLUSION

A presented CPSO algorithm technique is used to solve the Economic Dispatch (ED) problems. The CPSO algorithm had an adjustment in a position update strategy is added in the PSO framework to achieve the solutions and satisfying the constraints (both equality and inequality). With the help of multi dimensional search space, the CPSO technique is increase the convergence speed, and also the high probability constraints of ED problems cost function is satisfied with the help of global solution value. The taken benchmark sample systems are used to compare the existing algorithms with the presented proposed algorithms. These test systems are 6-units, 20-units, 38-units system.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Lee, Kwang Y, Mohamed A. [2008] El-Sharkawi, eds. Modern heuristic optimization techniques: theory and applications to power systems. Vol. 39. John Wiley & Sons

[2] Bakirtzis, Anastasios, Vassilios Petridis, and Spyros Kazarlis. [1994]Genetic algorithm solution to the economic dispatch problem. IEE proceedings-generation, transmission and distribution 141(4): 377-382.

[3] Lee, Fred N, Arthur M Breipohl.[1993] Reserve constrained economic dispatch with prohibited operating zones. IEEE transactions on power systems 8(1): 246-254.

[4] Yoshida, Hirotaka, et al. [2000]A particle swarm optimization for reactive power and voltage control considering voltage security assessment.IEEE Transactions on power systems 15(4): 1232-1239.

[5] Naka Shigenori, et al. [2001] Practical distribution state estimation using hybrid particle swarm optimization. Power Engineering Society Winter Meeting, 2001. IEEE. Vol. 2. IEEE.

[6] Wood Allen J, Bruce F Wollenberg. [2012] Power generation, operation, and control. John Wiley & Sons

[7] Kennedy James.[2011] Particle swarm optimization."Encyclopedia of machine learning. Springer US, 760-766.

[8] Yoshida, Hirotaka, et al.[2000] A particle swarm optimization for reactive power and voltage control considering voltage security assessment. IEEE Transactions on power systems 15(4): 1232-1239.

[9] Clerc Maurice, James Kennedy.[2002]The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE transactions on Evolutionary Computation 6(1): 58-73.

[10] Shi Yuhui, Russell Eberhart. [1998]Parameter selection in particle swarm optimization." Evolutionary programming VII. Springer Berlin/Heidelberg.

[11] Shi, Yuhui, Russell C. [1999] Eberhart. "Empirical study of particle swarm optimization." Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on. Vol. 3. IEEE.

[12] Bakirtzis, Anastasios, Vassilios Petridis, and Spyros Kazarlis.[1994] Genetic algorithm solution to the economic dispatch problem. IEE proceedings-generation, transmission and distribution 141(4): 377-382.

[13] Walters, David C, Gerald B Sheble.[1993] Genetic algorithm solution of economic dispatch with valve point loading. IEEE transactions on Power Systems 8(3): 1325-1332.

[14] Wong, Kit Po, Yin Wa Wong.[1994] Genetic and genetic/simulated-annealing approaches to economic dispatch. IEE Proceedings-Generation, Transmission and Distribution 141(5): 507-513.

[15] Sheblé, Gerald B, Kristin Brittig.[1995] Refined genetic algorithm-economic dispatch example." IEEE Transactions on Power Systems 10(1): 117-124.

[16] Walters, David C, Gerald B Sheble. [2005] Genetic algorithm solution of economic dispatch with valve point loading. IEEE transactions on Power Systems 8(3): 1325-1332.

[17] Chen, Po-Hung, and Hong-Chan Chang. [1995]Large-scale economic dispatch by genetic algorithm. IEEE transactions on power systems 10(4): 919-1926.

[18] Fung CC, Chow SY, Kit Po Wong. [2000] Solving the economic dispatch problem with an integrated parallel genetic algorithm." Power System Technology. Proceedings. PowerCon 2000. International Conference on. Vol. 3. IEEE,

[19] Yalcinoz T, Altun H, Uzam M. [2001] Economic dispatch solution using a genetic algorithm based on arithmetic crossover." Power tech proceedings, IEEE Porto. Vol. 2.

[20] Sun Shiyuan, Jianwei Li. [2014] A two-swarm cooperative particle swarms optimization. Swarm and Evolutionary Computation 15: 1-18.

[21] PANDIAN, SEVUGARATHINAM MUTHU VIJAYA, and Keppanagowder Thanushkodi.[2012] Considering transmission loss for an economic dispatch problem without valve-point loading using an EP-EPSO algorithm" TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES 20(2): 1259-1267.

[22] Bhattacharya, Aniruddha, Pranab Kumar Chattopadhyay.[2010] Hybrid differential evolution with biogeography-based optimization for solution of economic load dispatch. IEEE transactions on power systems 25(4): 1955-1964.

[23] Lachogiannis V, John G, Kwang Y Lee. [2009]Economic load dispatch—A comparative study on heuristic optimization techniques with an improved coordinated aggregation-based PSO." IEEE Transactions on Power Systems 24(2): 991-1001.
[24] Shaw Binod, et al. Solution of economic load dispatch problems by a novel seeker optimization algorithm. International Journal on Electrical Engineering and Informatics 3(1): 26.
[25] Bhattacharya, Aniruddha, Pranab Kumar Chattopadhyay. [2010]Biogeography-based optimization for different economic load dispatch problems. IEEE transactions on power systems 25(2):1064-1077.
[26] Gaing, Zwe-Lee.[2003] Particle swarm optimization to solving the economic dispatch considering the generator constraints. IEEE transactions on power systems 18(3): 1187-1195.
[27] Park, Jong-Bae, et al. A particle swarm optimization for economic dispatch with nonsmooth cost functions.IEEE Transactions on Power systems 20(1): 34-42.
[28] Pothiya, Saravuth, Issarachai Ngamroo, Waree Kongprawechnon.[2008] Application of multiple tabu search algorithm to solve dynamic economic dispatch considering generator constraints. Energy Conversion and Management 49(4): 506-516.
[29] Elsayed WT, et al. [2016]Modified social spider algorithm for solving the economic dispatch problem." Engineering Science and Technology, an International Journal 19(4): 1672-1681.
[30] Gaing, Zwe-Lee.[2003] Particle swarm optimization to solving the economic dispatch considering the generator constraints. IEEE transactions on power systems 18(3): 1187-1195.
[31] Selvakumar, Immanuel A, Thanushkodi K.[2007] A new particle swarm optimization solution to nonconvex economic dispatch problems. IEEE transactions on power systems 22(1): 42-51.
[32] Anusha B, Noah C Sivaranjani, Priyanka S. [2015]Predictive analysis of movie reviews using hybrid approach", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 1(1): 1-7.
[33] Dharshini G, Subhasri V, Sujitha G, Ganesan M, [2016]"Secure Information Retrival for Decentralised Disruption Tolerant Military Networks using CP-ABE", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(1): 1-6.
[34] Govindharaj I, Karthiga S, Manishalakshmi R, Mary Silvia Theodore R.[2016] Home Power Analyzer with Smart Power Monitoring using IoT", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(1): 7-13.
[35] Ahilandeswari T, Nandhini S, Sivasankari P, Rajalakshmy M, [2016] Intensifying the Generic Middleware for Smart Environment", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907,2(2): 1-5.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

OPPOSITIONAL CUCKOO SEARCH FOR SOLVING ECONOMIC POWER DISPATCH

Kalaipriyan Thirugnanasambandam[1*], J. Amudhavel[2], Sujatha Pothula[3]
[1,3]*Department of Computer Science, Pondicherry University, Puducherry, INDIA*
[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT
This paper proposes an efficient oppositional Cuckoo Search Algorithm (OCSA) for solving Economic Power Dispatch (EPD). This variant of Cuckoo Search (CS) is proposed in order to impose exploitation strategy in standard CS. Oppositional based Learning is a heuristic method which induces the intensification process of neighborhood solutions in the given search space. On the other hand, CS lacks in intensification phase since BSRW intuits more on exploration. Hence in this paper, OCSA in proposed to improve the intensification in non-smooth and non-linear EPD solution space. The proposed algorithm has been evaluated in with three different test systems in order to rove the efficiency of OCSA. For validating the performance of OCSA, different variants of CS and some of the novel evolutionary algorithms are tested in same simulation environment. The results show that the proposed OCSA outperforms existing algorithms and other variants of CS.

KEY WORDS
Oppositional Cuckoo search, economic power dispatch problem, non-linear search space, exploitation strategy, Opposition based learning.

Received: 11June 2017
Accepted: 20 July 2017
Published: 14 Sept 2017

***Corresponding Author**
Email: kalaip27@gmail.com
Tel: +91 9042208508

INTRODUCTION

In the field of power systems, EPD is one among the major concern since the need of electric energy in day today life become an essential and its growth towards the need gets increased in exponential manner. Distribution of power load to generation units in an economic way leads the problem to get solved using optimization algorithms. Without the constraints of EPD it seems to be a linear problem where the problem can be solved in polynomial time. But with the constraints like transmission loss total power demand, valve point effects, EPD turns out to be a non-linear optimization problem [1]. There are a number of variants available for distributing the power to generation units which are classified based on its mathematical formulation. Along with the EPD, emission of fossil fuels based generation has been coined as a multi-objective variant called combined economic emission dispatch problem (CEED) [2]. While solving EPD, dynamic models become a factor to be considered with ramp rate limits over 24 hours' horizon [3]. On incorporating renewable energy resources into existing EPD a probabilistic or stochastic formulation becomes essential [4]. In power systems, efficiency with respect to energy conversion can be achieve with the combined approach of heat and power namely Combined Heat and Power Economic Dispatch (CHPED) [5, 6]. Another variant, nonlinear EPD problem has been formulated which imposes ramp limits, valve point effects, prohibition in operation zones and different type of fuels with it [7, 1]. Conventional EPD used quadratic function to compute the total consumed fuel cost of generation units; however, in recent years since some of the generators holds steam valves opening processes sinusoidal points were added to conventional EPD for effective fuel cost consumption function [7]. In some other generators there exists limitation in fixing power ranges due to practical infeasibilities. This addressed issue has been coined as prohibited power zones and turned the search space into disjoint and non-smooth [1].

Researchers contributed numerous techniques to solve EPD and its variants using exact methods such as dynamic programming, linear programming, and so on. These methodologies are time consuming process since the time complexity of the algorithm increases exponentially as the number of generation units increased. Some of the heuristic methods are proposed including lambda iteration method, interior point method for solving conventional EPD problems. These heuristics fails to find best feasible solution when the search space becomes nonlinear. On the other hand, evolutionary algorithms have the trend of finding optimal feasible solution sets in non-linear search space. Genetic Algorithm [8], differential evolution [9] along with swarm based algorithms such as Particle Swarm Optimization [10], Seeker Optimization (SOA) [11], Harmony Search (HS) [12], Artificial Bee Colony (ABC) [13], Group Search Optimizer (GSO) [14], are some of the algorithms by which EPD has been solved. Recently, Zhou, et al [15] solved economic emission dispatch with respect to power security as an objective using Ant Colony Optimization (ACO). Imposing niche search phase in ACO promoted the individuals to result in efficient Pareto front solutions. Jiang, et al [16] addressed stochastic EPD in which wind integrated power system has been considered with demand response. Another proposed work on addressing wind integrated power system by Wang, et al [17] which has been solved using efficient heuristic method. Wei, et al. [18] solved environmental economic dispatch with wind and carbon capture plants using golden selection search algorithm. Qu, et al [19] solved EED using multi-objective evolutionary algorithm which addresses wind generated power system. Shilaja and Ravi [20] addressed EPD problem with solar power plants as power units and tested their approach in standard IEEE 30 and 57 bus systems.

Based on the research contributions made in EPD and its variants, in this approach an attempt to solve EPD using a variant of Cuckoo Search (CS) namely Oppositional Cuckoo Search Algorithm (OCSA) has been made. This method concentrates on intensification phase in standard cuckoo search algorithm where a consistent exploration and exploitation phase is followed throughout the search. CS has been chosen since it is a lightweight algorithm (in terms of memory) [21-26].

IIOAB JOURNAL
ISSN: 0976-3104

The reminder of this paper is as follows: Section 2 holds the definition of EPD along with the constraints considered, Section 3 addresses standard CS and OCSA, Section 4 holds the experimental evaluation and result discussion on three different Test systems, Section 5 holds the conclusion and future work.

EPD PROBLEM FORMULATION

This problem formulation considers EPD as a minimization problem whose objective is to minimize the total generation cost with satisfied equality (power demand) and inequality constraints (power consumption).

Objective formulation

As stated above the objective is to minimize the total fuel cost consumed by the generators. The cost function for i^{th} generation unit can be formulated as follows:

$$F_i = a_i P_i^2 + b_i P_i + c_i$$

where a, b, c are the cost coefficients of i^{th} generator unit.
On imposing valve point effects on the cost functions which is a variant formulation of EPD, the search space becomes non-smooth and non-linear. The mathematical formulation of EPD with valve point effects can be represented as

$$F_i = a_i P_i^2 + b_i P_i + c_i + \left| e_i \sin(f_i (P_i^{min} - P_i)) \right|$$

where e and f are the coefficients of valve point effects, P^{min} is the lower bound value of P. The bound of P for each generator unit are not same.
Thus, the objective of EPD with valve point effects can be formulated as

$$Minimize\ F_c = \sum_{i=1}^{n} a_i P_i^2 + b_i P_i + c_i + \left| e_i \sin(f_i (P_i^{min} - P_i)) \right|$$

where n represents the total number of generators
Constraints
The equality and inequality constraints to be satisfied in economic dispatch problem in terms of generating capacity and power balance are as follows.

Power balance

The total generated power is the summation of the total power demand and transmission loss in the power system. The constraint is mathematically formulated as.

$$\sum_{i=1}^{n} P_i = P_D + P_{TL}$$

where P_D is the total power demand in the system and P_{TL} is the total transmission loss in the system.
The total transmission loss is calculated using B coefficient and the formulated is represented as follows.

$$P_{TL} = \sum_{i=1}^{n} \sum_{j=1}^{n} P_i B_{ij} P_j + \sum_{i=1}^{n} B_{0i} P_i + B_{00}$$

where B_{ij}, B_{0i} and B_{00} are the i,jth loss coefficient of the symmetric matrix B, ith element of the loss coefficient vector and loss coefficient constant respectively.

Generating capacity constraint

The power output of each generation must satisfy generating capacity constraints and it is represented as.

$$P_i^{min} \leq P \leq P_i^{max}$$

where P_i^{min} and P_i^{max} are the minimum and maximum limits of the power output of generator i.

CUCKOO SEARCH ALGORITHM

Yang and Deb developed CS based on the inspiration from cuckoo's brood parasitism [27]. CS is simple but yet an effective algorithm which follows Markov chain model (next generation population is based on the current population). Memory-wise, CS is considered to be a powerful lightweight algorithm since there will not be any storage of previous best solutions and global best solutions. In CS, the solutions are generated based on two different strategies Levy Flights Random Walk (LFRW) and Biased/ Selective Random Walk (BSRW).

In a given search space X where $x_{d,min}$ and $x_{d,max}$ are the lower and upper bounds of search space where $d = [1,2,\ldots D]$ represents the dimensions of search space. An individual in CS can be represented as $Indv = [x_1, x_2, \ldots, x_D]$

and the population of CS can be denoted as

$$pop = [Indv_1, Indv_2, \dots, Indv_N]$$

where N represents total number of individuals in the population.
Each individual in CS can be generated randomly as

$$Indv = x_{d,min} + rand() \times (x_{d,max} - x_{d,min}) \quad (1)$$

where $d = [1,2,\dots D]$. After the process of initialization CS goes for LFRW method for searching process in the given search space.
LFRW imposes exploration phase of search in the given search space with the help of Levy Flights whose concentration of randomization can be limited with step size α. Based on LFRW the generation of random solutions are as follows:

$$Y_i = Indv_i + \alpha \oplus Levy\ (\beta) \quad (2)$$

where Y_i is the newly generated solution of $Indv_i$. $Levy\ (\beta)$ can be represented using the formula of power-law

$$Levy\ (\beta) = r^{-1-\beta} \quad (3)$$

where r is the random variable, β ranges (0,2) termed as stability factor During implementation process, $Levy\ (\beta)$ is followed as [28]:

$$Levy\ (\beta) = \frac{\emptyset \times \mu}{|v|^{\frac{1}{\beta}}} \quad (4)$$

where μ and v are the random variables which ranges (0,1) and the value of $\emptyset$ can be calculated as

$$\emptyset = \left[\frac{\sin\left(\frac{\pi \times \beta}{2}\right) \times \Gamma(1+\beta)}{2^{\left(\frac{\beta-1}{2}\right)} \times \Gamma\left(\frac{1+\beta}{2}\right) \times \beta} \right]^{\frac{1}{\beta}} \quad (5)$$

where Γ is $gamma\ function.$
After the process of LFRW, CS imposes the obtained solutions to BSRW where the search has been done based on greedy method. In this search a probability factor P_a has been defined which limits the total number of individuals by deleting the worst individuals below P_a.
After the worst solutions are abandoned a new set of solutions will be generated as follows:

$$Z_i = Indv_i + rand \times (Indv_g - Indv_h) \quad (6)$$

where $i \in |abandon\ solutions|$, $rand$ denotes a random variable range (0,1), $Indv_g$, $Indv_h$ represents randomly generated solutions based on Eq. (1).
In standard CS, since the Levy flight is controlled by step size there exists a consistent phase between exploration and exploitation. This leads CS to divergence from optimal solution. Hence the authors of this paper proposed a concept with high exploitation strategy in the given search space which is an essential factor in EPD optimization problem. Pseudo code for standard CS is given in Algorithm 1.

Cuckoo search algorithm:

Input: Population $pop = [Indv_1, Indv_2, \dots, Indv_N]$
Each $Indv$ consists of dimensions ($1,2,\dots,D$)
Pop, Termination Criteria, α, P_a, min, max
1: Initialize the population pop
2: for $i = 1\ to\ Pop_{Next}$
3: $Indv_i = (max - min) * rand + min$
4: Compute $f(Indv_i)$
5: end
6: while (Termination Criteria not satisfied) do
7: Generate random individuals Y_i based on Eq. (2) where $i \in 1 \dots N$
8: Compute $f(Y_i)$ where $i \in 1 \dots N$
9: for $i = 1\ to\ N$
10: if ($f(Y_i) \leq f(Indv_i)$
11: $Indv_i = Y_i$
12: $f(Indv_i) = f(Y_i)$
13: end
14: end
15: for $i = 1\ to\ N$
16: if (rand() < P_a)
17: $Indv_i =$ Generate Z_i based Eq. (6)

18: end
19: end
20: $f(Indv_{best}) = \min(f(Indv_i))$
21: Track $Indv_{best}$
22: end while
Output: $Indv_{best}$ and $f(Indv_{best})$

Algorithm 1: Pseudo code of standard Cuckoo Search Algorithm.

OPPOSITIONAL BASED LEARNING

Tizhoosh [28] proposed a phase called Oppositional Based Learning (OBL) for improving the convergence speed of optimization techniques towards optimal solution. This concept produces an opposition of current individual and evaluates the performance of the generated individual and current individual based on the jumping rate. This process finds better individual in the given search space to provide optimal solution at the end of initiated search. OBL has been successfully used in more number of meta-heuristics [30, 31] which enhances the convergence speed towards optimal solution. To implement OBL, opposite number of OBL has to be defined.

When N such that $N \epsilon [a, b]$ be a real number where a and b are the upper and lower bounds of N the opposite number N^0 can be defined as

$$N^0 = a + b - N$$

When oppositional concept implemented on more than one-dimension problem it has to be formulated as

$$N_d^0 = a_d + b_d - N_d$$

where $i \epsilon 1 \leq d \leq D$, D represents the number of dimensions, $N_d \epsilon [a_d, b_d]$.

OPPOSITIONAL CUCKOO SEARCH

In OCSA, the standard CS is incorporated with Oppositional based learning in BSRW. When the individuals are generated with BSRW, the concept of extracting information from two different random individuals are replaced with oppositional based learning concept. Thus Eq. (6) can be redefined as

$$Z_{i,d} = Indv_{i,d} + rand \times N_{i,d}^0$$

The pseudo code of proposed OCSA is given in Algorithm 2.

Oppositional Cuckoo search algorithm (OCSA)

Input: Population $pop = [Indv_1, Indv_2, \ldots, Indv_N]$
Each $Indv$ consists of dimensions $(1, 2, \ldots, D)$
Pop, Termination Criteria, a, P_a, min, max
1: Initialize the population pop
2: for $i = 1 \; to \; Pop_{Nest}$
3: $Indv_i = (max - min) *. rand + min$
4: Compute $f(Indv_i)$
5: end
6: while (Termination Criteria not satisfied) do
7: Generate random individuals Y_i based on Eq. (2) where $i \in 1 \ldots N$
8: Compute $f(Y_i)$ where $i \in 1 \ldots N$
9: for $i = 1 \; to \; N$
10: if ($f(Y_i) \leq f(Indv_i)$
11: $Indv_i = Y_i$
12: $f(Indv_i) = f(Y_i)$
13: end
14: end
15: for $i = 1 \; to \; N$
16: if (rand() < P_a)
17: $Indv_i =$ Generate Z_i based Eq. (7)
18: end
19: end
20: $f(Indv_{best}) = \min(f(Indv_i))$

21: Track $Indv_{best}$
22: end while
Output: $Indv_{best}$ and $f(Indv_{best})$

Algorithm 2: Pseudo code of OCSA.

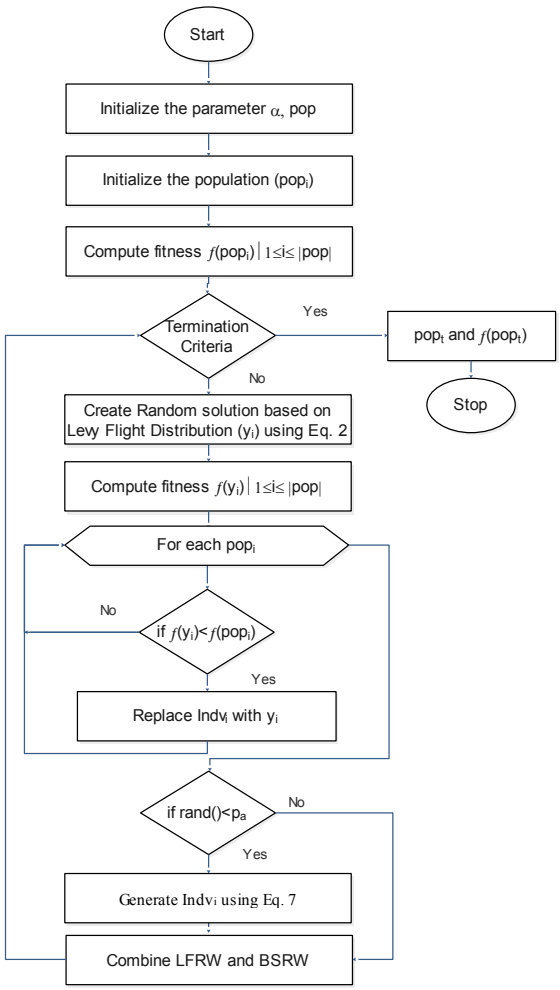


Figure 1: Flow chart of oppositional cuckoo search.

Experimental Evaluations and Result Discussion

To evaluate the performance of the proposed algorithm on EPD, three different power systems which holds 10 power units [31] 13 power units [32] and 40 power units [33] are considered. OSCA and other compared algorithms were implemented in MATLAB 8.3 with the system configuration of Intel core i7 processor with 3.2 GHz speed and 4GB RAM. Parameter settings for experimental results are tabulated in [Table 1].

Table 1: Parameter settings for Experimental Evaluation

Type	Method
Total Individuals	100
Maximum Iterations	1000
Heuristic Used	Oppositional based Learning
Termination Condition	Maximum Iterations
Run	20
α	0.1

Test System 1

In this test system 10 power units are considered for effective power transmission to generation units. The system has been evaluated with power demand 500 MW. The system considers valve point effects along with power balance and transmission loss. Table II holds the simulation results of 10-unit power

systems with a power demand of 500MW. To evaluate the performance of the proposed algorithm on EPD, the results are compared with existing techniques such as GGCS [34], NNCS [35], HSACS [36], PSO [37], Gradient Search [38] and CECS [39]. On comparing the results of proposed algorithm in the simulated environment, the results show that the proposed algorithm outperforms existing techniques in terms of fuel cost consumption. The test data are found in [31]. Evaluation of the performance of OCSA on another case of 10-unit power system which consists of 500MW were done with same test data [31] and the results are tabulated in [Table 2].

Table 2: Unit output of different methods for test case 1

Unit	GGCS	OCSA	NNCS	HSACS	CECS	PSO	Gradient search
P1	23.60743	88.20742	24.37497	14.03441	18.90339	12.3000	13.4868
P2	23.07166	13	13.09404	13.16896	13	14.4134	13.5586
P3	10.83128	10	11.69235	24.18451	15.86997	10.0069	12.2281
P4	37.34063	25.90333	33.39361	15	16.82842	26.3741	18.3283
P5	54.80545	54.54007	89.31385	72.00139	72.24111	87.5469	125
P6	14.69713	28.30135	53.91961	33.17024	44.00295	50.5875	20.6141
P7	46.45713	34.71568	20	55.5201	30.1395	86.3165	93.8312
P8	25.61695	25	25	25	33.70554	28.0985	25
P9	150	150	150	150	150	75.9916	34.5543
P10	126.7559	144.8579	92.91588	99.89749	113.9942	110.76	150
Total Loss	13.1836	74.52572	13.70432	1.977099	8.685112	2.21323	2.11552
Total Cost	11069.75	**10948.83**	11096.99	11076.72	11105.97	11080.5	13340.5

Test System 2

In this test system 13 power units are considered for effective power transmission to generation units. The system has been evaluated with the power demand 2520 MW. The system considers valve point effects along with power balance and transmission loss. [Table 3] holds the simulation results of 13-unit power systems with a power demand of 2520MW. To evaluate the performance of the proposed algorithm on EPD, the results are compared with existing techniques such as OGOW [40], GWO [41], OIWO [42], SDE [43] and ORCCRO [44]. On comparing the results of proposed algorithm in the given test bed, the results show that the proposed algorithm outperforms existing techniques in terms of fuel cost consumption. The test data are found in [32].

Table 3: Unit output of different methods for test case 2

Unit	OGWO	GWO	OIWO	SDE	ORCCRO	OCSA
P1	628.2940	628.1678	628.3185	628.32	628.32	549.5416
P2	299.1803	298.9229	299.1989	299.20	299.20	278.7213
P3	297.5041	298.2269	299.1991	299.20	299.20	359.9482
P4	159.7284	159.7232	159.7331	159.73	159.73	148.0571
P5	159.7325	159.7210	159.7331	159.73	159.73	173.2179
P6	159.7295	159.7270	159.7331	159.73	159.73	179.0255
P7	159.7334	159.7173	159.7330	159.73	159.73	174.535
P8	159.7323	159.6793	159.7331	159.73	159.73	148.8514
P9	159.7327	159.6673	159.7330	77.40	77.40	142.4425
P10	77.3963	77.3971	77.3953	113.12	112.14	66.26028
P11	114.7487	114.6051	113.1079	92.40	92.40	107.6776
P12	92.3974	92.3886	92.3594	92.40	92.40	97.00774
P13	92.3780	92.3550	92.3911	92.40	92.40	91
Total Loss	40.2874	40.2983	40.3686	40.43	39.43	16.286
Total Cost	24512.7250	24514.4774	24514.83	24514.88	24513.91	24320.35

Test System 3

In this test system 40 power units are considered for evaluating the performance of proposed algorithm in the simulated environment. The system has been evaluated with the power demand 10500 MW. Transmission loss is neglected in 40 generation power systems for comparison purpose of the proposed algorithm. [Table 4] holds the simulation results of 40-unit power systems with a power demand of 10500 MW. To validate the performance of the proposed algorithm on EPD, the results are compared with other variants of Cuckoo search such as [34], NNCS [35], HSACS [36] and CECS [39]. On comparing the results of proposed algorithm in the given test bed, the results show that the proposed algorithm outperforms existing techniques in terms of fuel cost consumption. The test data are found in [33].

Table 4: Unit output of different methods for test case 3

Unit	GGCS	NNCS	HSACS	GGCS	CECS	OCSA
P1	50.70948	64.08755	74.42451	36	91.39232	36
P2	114	110.7719	110.7608	113.0483	105.9526	101.9321
P3	120	120	66.37267	71.69646	114.2243	82.0777
P4	190	160.4973	168.8296	146.588	189.7303	168.7644
P5	97	87.13556	97	97	97	97
P6	105.5822	129.7955	137.3084	91.71777	104.5073	84.9458
P7	208.5234	219.6922	300	298.7696	230.0757	297.9823

P8	290.6988	268.0003	291.8245	268.5739	243.3768	290.2042
P9	166.1162	209.1052	177.0825	270.1083	274.1762	245.0909
P10	214.1494	288.5109	182.6113	142.4041	294.3056	182.892
P11	375	299.4434	333.2223	335.1445	159.5352	156.1595
P12	268.3552	328.7383	193.0712	341.1426	94	260.9524
P13	500	329.5528	267.5022	220.3067	368.5003	454.9511
P14	195.7186	370.5926	500	500	195.2312	412.1025
P15	215.1643	331.1107	500	193.1617	376.9174	240.4924
P16	198.575	500	436.8967	438.3321	500	358.4201
P17	500	452.649	499.5347	457.9635	404.5655	500
P18	500	418.0952	498.5542	492.5126	481.227	500
P19	547.7405	318.0209	517.5683	406.1257	495.5848	526.2942
P20	550	462.2571	270.1213	548.2812	522.3797	550
P21	550	549.6432	396.4421	506.9649	485.074	550
P22	447.5882	549.5902	442.9756	260.5514	532.9034	539.2501
P23	550	301.8208	535.9142	528.2371	530.896	550
P24	550	546.6685	549.9532	534.3838	492.7228	333.6466
P25	543.1754	506.7675	533.5099	550	395.0965	549.994
P26	550	548.5131	466.9563	545.9104	544.8522	550
P27	38.9467	38.97291	21.4302	10	29.29106	10
P28	10	42.99519	10	18.56083	49.74159	10
P29	25.19344	17.44346	15.56426	46.20762	36.6943	18.60118
P30	97	97	97	97	97	87.98265
P31	160.2971	190	190	190	190	190
P32	190	190	190	190	190	190
P33	190	190	190	190	190	176.3939
P34	181.8816	115.2958	108.9326	200	187.0463	108.4217
P35	142.4572	200	199.1489	124.8424	182.3224	90
P36	200	156.5702	127.998	176.376	200	90
P37	100.932	53.14813	76.0706	76.09422	76.68901	110
P38	105.0013	72.60135	89.30453	79.88293	106.1308	103.1142
P39	52.85066	104.2112	71.91103	78.7876	109.8352	103.4303
P40	345.9452	510.897	524.6997	549.7684	508.5204	515.4442
TOTAL COST	124053.7	125878.4	124684.6	124885.8	123947.2	**123189.6**

CONCLUSION

In this work, an effective OCSA is proposed for solving EPD with valve point effects, equality and non-equality constraints. Three test cases are used in this paper for evaluating and validating the proposed algorithm which consists 10, 13 and 40 power systems. The advantage of oppositional based learning is to intensify the search towards exploitation for obtaining optimal solutions. Experimental results show that OCSA outperforms in terms of fuel cost consumption. The convergence of the proposed algorithm towards optimal solution is higher when compared with other compared algorithms. Future enhancement of this work can be done with more limit factors of EPD.

CONFLICT OF INTEREST
There is no conflict of interest

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Ding T, Bo R, Li F, Sun H. [2015] A bi-level branch and bound method for economic dispatch with disjoint prohibited zones considering network losses, IEEE Transactions on Power Systems, 30(6):2841-2855.

[2] Jadoun VK, Gupta N, Niazi KR, Swarnkar A. [2015] Modulated particle swarm optimization for economic emission dispatch. International Journal of Electrical Power & Energy Systems, 73: 80-88.

[3] Elattar, Ehab E. [2015] A hybrid genetic algorithm and bacterial foraging approach for dynamic economic dispatch problem. International Journal of Electrical Power & Energy Systems 69: 18-26.

[4] Dubey, Hari Mohan, Manjaree Pandit, Panigrahi BK. [2015] Hybrid flower pollination algorithm with time-varying fuzzy selection mechanism for wind integrated multi-objective dynamic economic dispatch. Renewable Energy 83: 188-202.

[5] Jayakumar N, Subramanian S, Ganesan S, Elanchezhian EB. [2016] Grey wolf optimization for combined heat and power dispatch with cogeneration systems. International Journal of Electrical Power & Energy Systems, 74:252-264.

[6] Beigvand, Soheil Derafshi, Hamdi Abdi, Massimo La Scala. [2016] Combined heat and power economic dispatch problem using gravitational

search algorithm." Electric Power Systems Research 133: 160-172.
[7] Elsayed, Wael T, Ehab F El-Saadany. [2015] A fully decentralized approach for solving the economic dispatch problem. IEEE Transactions on Power Systems 30(4): 2179-2189.
[8] Walters, David C, Gerald B Sheble. [1993] Genetic algorithm solution of economic dispatch with valve point loading. IEEE transactions on Power Systems 8(3): 1325-1332.
[9] Gaing, Zwe-Lee. [2003] Particle swarm optimization to solving the economic dispatch considering the generator constraints. IEEE transactions on power systems 18(3): 1187-1195.
[10] Noman, Nasimul, Hitoshi Iba. [2008] Differential evolution for economic load dispatch problems. Electric Power Systems Research 78(8): 1322-1331.
[11] Shaw Bikash, V Mukherjee, Sakti Prasad Ghoshal. [2011] Seeker optimisation algorithm: application to the solution of economic load dispatch problems. IET generation, transmission & distribution 5(1): 81-91.
[12] Wang, Ling, and Ling-po Li. [2013] An effective differential harmony search algorithm for the solving non-convex economic load dispatch problems. International Journal of Electrical Power & Energy Systems 44(1): 832-843.
[13] Secui, Dinu Calin. [2015] The chaotic global best artificial bee colony algorithm for the multi-area economic/emission dispatch. Energy 93: 2518-2545.
[14] Moradi-Dalvand M, Mohammadi-Ivatloo B, Najafi A, Rabiee A. [2012] Continuous quick group search optimizer for solving non-convex economic dispatch problems. Electric Power Systems Research, 93: 93-105.
[15] Zhou J, Wang C, Li Y, Wang P, Li C, Lu P, Mo L. [2017] A multi-objective multi-population ant colony optimization for economic emission dispatch considering power system security. Applied Mathematical Modelling.
[16] Jiang Y, Xu J, Sun Y, Wei C, Wang J, Ke D, Tang B. [2017] Day-ahead stochastic economic dispatch of wind integrated power system considering demand response of residential hybrid energy system. Applied Energy, 190: 1126-1137.
[17] Wang, Xu, Chuanwen Jiang, Bosong Li. [2016] Active robust optimization for wind integrated power system economic dispatch considering hourly demand response. Renewable Energy 97: 798-808.
[18] Wei W, Liu F, Wang J, Chen L, Mei S, Yuan T. [2016] Robust environmental-economic dispatch incorporating wind power generation and carbon capture plants, Applied Energy, 183:674-684.
[19] Qu BY, Lian, JJ, Zhu YS, Wang ZY, Suganthan PN. [2016] Economic emission dispatch problems with stochastic wind power using summation based multi-objective evolutionary algorithm. Information Sciences, 351: 48-66.
[20] Shilaja C, Ravi K. [2017] Optimization of emission/economic dispatch using euclidean affine flower pollination algorithm (eFPA) and binary FPA (BFPA) in solar photo voltaic generation. Renewable Energy 107: 550-566.
[21] Nguyen, Thang Trung, Dieu Ngoc Vo, Bach Hoang Dinh. [2016] Cuckoo search algorithm for combined heat and power economic dispatch." International Journal of Electrical Power & Energy Systems 81: 204-214.
[22] Nguyen, Thang Trung, and Dieu Ngoc Vo. [2015] The application of one rank cuckoo search algorithm for solving economic load dispatch problems. Applied Soft Computing 37: 763-773.
[23] Mellal, Mohamed Arezki, Edward J Williams. [2015] Cuckoo optimization algorithm with penalty function for combined heat and power economic dispatch problem. Energy 93: 1711-1718.
[24] Sekhar, Pudi, and Sanjeeb Mohanty. [2016] An enhanced cuckoo search algorithm based contingency constrained economic load dispatch for security enhancement. International Journal of Electrical Power & Energy Systems 75: 303-310.
[25] Basu, M., and A. Chowdhury. [2013] Cuckoo search algorithm for economic dispatch. Energy 60: 99-108.
[26] Nguyen, Thang Trung, and Dieu Ngoc Vo. [2016] An efficient cuckoo bird inspired meta-heuristic algorithm for short-term combined economic emission hydrothermal scheduling. Ain Shams Engineering Journal.
[27] Yang, Xin-She, and Suash Deb. [2010] Engineering optimization by cuckoo search. International Journal of Mathematical Modelling and Numerical Optimization 1.4: 330-343.
[28] Tizhoosh, Hamid R. [2015] Opposition-based learning: a new scheme for machine intelligence. Computational intelligence for modelling, control and automation, 2005 and international conference on intelligent agents, web technologies and internet commerce, international conference on. Vol. 1. IEEE.
[29] Roy, Provas Kumar, Chandan Paul, and Sneha Sultana. [2014] Oppositional teaching learning based optimization approach for combined heat and power dispatch. International Journal of Electrical Power & Energy Systems 57: 392-403.
[30] Roy, Provas Kumar, and Dharmadas Mandal. [2014] Oppositional biogeography-based optimisation for optimal power flow." International Journal of Power and Energy Conversion 5(1): 47-69.
[31] Sen, Tanuj, and Hitesh Datt Mathur. [2016] A new approach to solve Economic Dispatch problem using a Hybrid ACO–ABC–HS optimization algorithm." International Journal of Electrical Power & Energy Systems 78: 735-744.
[32] Črepinšek, Matej, Shih-Hsi Liu, Marjan Mernik. [2013] Exploration and exploitation in evolutionary algorithms: A survey.ACM Computing Surveys (CSUR), 45(3):3.
[33] Nidul Sinha,, Chakrabarti R, PK. Chattopadhyay. [2003] Evolutionary programming techniques for economic load dispatch. IEEE Transactions on evolutionary computation 7(1): 83-94.
[34] Dhabal, Supriya, Palaniandavar Venkateswaran. [2017] An efficient gbest-guided Cuckoo Search algorithm for higher order two channel filter bank design. Swarm and Evolutionary Computation 33: 68-84.
[35] Wang, Lijin, Yiwen Zhong, Yilong Yin. [2016] Nearest neighbour cuckoo search algorithm with probabilistic mutation. Applied Soft Computing 49: 498-509.
[36] Mlakar Uroš, Iztok Fister.[2016] Hybrid self-adaptive cuckoo search for global optimization. Swarm and Evolutionary Computation 29: 47-72.

[37] Govindharaj I, Karthiga S, Manishalakshmi R, Mary Silvia Theodore R. [2016] Home Power Analyzer with Smart Power Monitoring using IoT", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(1): 7-13.
[38] Rajadurai R, Amelia, Aubrey, Anusha A, Danapriya P , Geethashnee D. [2017] Efficient Data Leakage Prevention Strategy using Key Distribution", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 3(1): 9-16.
[39] Anusha B, Noah C. Sivaranjani, Priyanka S, [2015] Predictive analysis of movie reviews using hybrid approach", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907,1(1): 1-7.
[40] Pradhan, Moumita, Provas Kumar Roy, and Tandra Pal. [2017] Oppositional based grey wolf optimization algorithm for economic dispatch problem of power system. Ain Shams Engineering Journal.
[41] Wong, Lo Ing, MH Sulaiman, Mohamed MR, Mee Song Hong. [2014] Grey Wolf Optimizer for solving economic dispatch problems.In Power and Energy (PECon) IEEE International Conference on, pp. 150-154. IEEE.
[42] Bhattacharjee, Kuntal, Aniruddha Bhattacharya, and Sunita Halder nee Dey. [2014] Oppositional real coded chemical reaction optimization for different economic dispatch problems. International Journal of Electrical Power & Energy Systems 55: 378-391.
[43] Ciornei, Irina, and Elias Kyriakides. [2012]A GA-API solution for the economic dispatch of generation in power system operation. IEEE Transactions on power systems 27(1): 233-242.
[44] Bhattacharya, Aniruddha, Pranab Kumar Chattopadhyay. [2010] Biogeography-based optimization for different economic load dispatch problems. IEEE transactions on power systems 25(2): 1064-1077.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

ENHANCED ARTIFICIAL BEE COLONY OPTIMIZATION FOR SOLVING ECONOMIC LOAD DISPATCH

R.S. Raghav[1*], J. Amudhavel[2], P. Dhavachelvan[1]
[1] *Department of Computer Science, Pondicherry University, Pondicherry, INDIA*
[2] *Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT

The power system needs accurate solution for carrying Economic load dispatch (ELD) optimization problems. The constraints involved in ELD are complex and it requires a perfect algorithm to minimize the cost and reduce the transmission loss. There are certain methods for solving the issues but they fail to achieve the target by retained in local optima. In this paper we proposed an enhanced Artificial Bee Colony optimization algorithm with new search technique. It inherits the behaviour of honey bees, to solve ELD problem in an effective way. It increases the local exploitation capability and avoids the premature convergence. The effectiveness of EABC algorithm is verified by and its performance is validated using the four test cases units. The algorithm is compared with other methods and simulation results show that algorithm suppress the performance of other traditional approaches, by generating better results.

KEY WORDS
Economic load dispatch (ELD), Artificial Bee Colony Optimization (ABC), Transmission Loss, Economic Dispatch (ED)

Received: 27 June 2017
Accepted: 23 July 2017
Published: 14 Sept 2017

***Corresponding Author**
Email: vpmrags@gmail.com
Tel.: +91-9952328330

INTRODUCTION

The core attributes of Economic Dispatch (ED) for handling energy generation and distribution at present days is major task. The requirement of energy and the cost of the fuel get increased, it also reflects by maximizing the cost of the entire system. In Economic Dispatch load schedules are optimized for the generators to achieve the proper power system, and it also needs to provide complete power demand with minimum generating cost [1]. The result comprises of mathematical optimization techniques which explains about the cost function curves. The curves for power unit generator need to be formulated [6]. It is adapted to the system inorder to find the optimal allocation of the load. The main objective of the system in EDP is to attain minimal cost and to utilize the complete power demand.

Some of the evolutionary algorithms are noted such as Particle Swarm Optimization (PSO), Real-Coded Genetic Algorithm (RCGA), Differential Evolution (DE), and Covariance Matrix Adapted Evaluation strategy (CMAES). The above mentioned Traditional approaches consist of some issues, which don't have the ability to fulfil the requirements of the power output generating units. It faces many issues such as converging to the local optimum, consuming more cost etc. Hence we need to go for some complex algorithms which satisfy the requirements of the system [2]. The role of Swarm Intelligence is to provide optimal results for complex problem. In the paper, we discuss about the artificial bee colony optimization with enhanced exploration.

LITERATURE STUDY

Lu, Peng, et al. [1] discuss about the dynamic economic dispatch, the performance of the scheme in thermal system works efficient. Here valve point effect is considered for non-smooth and non-convex to handle the DED problem. In this paper chaotic differential bee colony optimization algorithm (CDBCO) is endorsed to resolve issues of premature convergence. It helps to improvise the local exploitation capacity by incorporating a chaotic local search (CLS) method. The proposed CDBO algorithm is showcased by using four test case units and the comparison of results. The simulated results for the proposed scheme outperform other results with less computational time. Gaing, Zwe-Lee [2] enunciates a particle swarm optimization (PSO) method for handling the economic dispatch(ED) problem. The ramp rate limits, Prohibited operating zone and non smooth cost function are non-linear attributes. The proposed method is demonstrated for three different systems and the comparison is carried with GA method. The experiments results displays that the proposed PSO method has the ability to provide better results. Zhou, Jianzhong, et al [3] discuss about environmental issues such as global warming increases everyday it has drawn more focused towards daily optimization of electric power systems. The aim of Economic emission dispatch (EED) is to deduct the pollution produced by power generation, and these are proposed as non- convex, multi-objective and non-linear optimization problem. In a functional power system, the problem of EED becomes more intrigue among the objectives of the economy and emission, valve point effect, prohibited operation zones of generating units, and security constraints of transmission networks. To resolve such intrigued problems, an algorithm of a multi-objective-population ant colony optimization for uninterrupted domain (MMACO_R)is proposed. MMACO_R redesigns the pheromone structure of the ant colony

IIOAB JOURNAL
ISSN: 0976-3104

to expand the primary uni-objective method to multi-objective area. Moreover, to intensify the searching ability and overcomes early convergence. Multi-population ant colony is also propounded, which consists of ant populations with varied searching scope and speed. Along with that, a Gaussian function based niche search method is proposed to increase efficiency in distribution and accuracy of solutions on the Pareto optimal front. To confirm the working of MMACO_R in various multi-objective problems, conventional test are conducted. And ultimately, the algorithm that was proposed is applied to solve the EED based on a six-unit system, a tent-unit system and a standard IEEE 30-bus system. Simulation Results displays that MMACO_R is efficient in solving economic emission dispatch in a functional power system. Basu, M[4] discuss about the multi-area economic dispatch using the artificial bee colony optimization. In this paper some of constraints are considered such as prohibited operating zones, loss occurs during transmission, limitations for multiple fuels and valve-point loading. Three variety of system is used to evaluate the performance of the proposed system with variety in degree of complexity. The outcomes are compared with differential evolution, evolutionary programming and real coded genetic algorithm by knowing some of the important factors. The results show the proposed algorithm provides efficient results for handling MAED problems in practical power system.

Jadhav and Roy [5] propose Gbest guided artificial bee colony algorithm (GABC) inorder to solve the complex constraints in the power system. Here they took wind thermal power system problem modelled with weibull probability distribution function (PDF). Moreover, optimizing the cost of wind power is critical task, where it requires special methods to solve the issue. To calculate the performance of the proposed method, it is implemented to three standard test systems. The constraint used in the first scenario is considering different technical constraints such as prohibited zones, ramp rate limits valve loading effect, etc. In second scenario IEEE-30 bus test system is sued for evaluation. The proposed optimization techniques reveal that the proposed technique has better solution accuracy and convergence results.

Sen, Tanuj, and Hitesh Datt Mathur [6] present a newly developed hybrid optimization algorithm for handling the Economic Dispatch (ED) issues for a multi-generator system. The combination of Ant Colony Optimization (ACO), Artificial Bee Colony (ABC) and Harmonic Search (HS) algorithms is merged to form a hybrid algorithm. This strategy helps to identify optimized solution for the system, by improvising each solution given by the ACO module. The role HS module is used to remove the low value from the solution set and replace the better ones. The efficiency of this hybrid algorithm is estimated with other conventional ED solving methods like Gradient Search, evolutionary algorithms. The constraints like valve points, transmission loss are considered to be the key factors for comparison. The results show better performance with minimize the cost and loss.

Abdelaziz, A. Y., E. S. Ali, and SM Abd Elazim[7] propose a method for minimizing the operation cost by mapping needed load between the obtainable generation units. The non linear based constrained optimization is problem is formulated with both equality and inequality constraints. The emission of gaseous pollutants of fossil-fuelled power plants is used by the dual-objective Combined Economic Emission Dispatch (CEED) problem. In this paper, handling ELD and CEED issues in power system is carried by incorporating Flower Pollination Algorithm (FPA). The outcome is compared with other optimization algorithms for various power systems. The generated results from the proposed algorithm outperform other techniques even for massive power system with less computational time. Secui, Dinu Calin [8] discuss about the random sequences used for solution updating of GBABC (global best artificial bee colony algorithm). By using chaotic maps generated by the chaotic sequence replaces chaotic optimization method to provide solution to the multi-area economic/emission dispatch problem. To solve the problem some constraints are imposed such as multi-fuel sources, tie line capacity, valve-point effects and transmission loss. To study the behaviour of the algorithm, ten chaotic maps are implemented by both one dimensional and bi-dimensional. The efficiency of CGBABC algorithms is tested on five systems (6-unit, 10-unit, 16-unit, 40- unit and 120-unit) with variety characteristics, sizes. They solved by incorporating different chaotic maps. The obtained results are compared with other chaotic maps and they display better performance than the classical ABC algorithm, the GBABC algorithm and other optimization techniques.

PROBLEM FORMULATION

The main theme of Economic Dispatch is to identify optimum load allotment to the generators in a power system. The total fuel cost is reduced to the nominal value and to reduce the transmission [9-11]. The total fuel cost linked with the power system for understanding the requirements of the system.

Objective function

The key objective for solving the Economic dispatch is to reduce total generation cost for acquiring load demand of the power system by satisfying variety of constraints [12]. The fuel cost can be achieved nominal by

using the quadratic function of output power generator. The quadratic cost function is enoted by using the below equation

$$F_i = a_i p_{gi}^2 + b_i p_{gi} + c_i \ldots\ldots (1)$$

$$\sum_{i=1}^{n} F_i(p_{gi}) = \sum_{i=1}^{n} [a_i (p_{gi})^2 + b_i (p_{gi}) + c_i] \text{ where i=1,2,...,n} \ldots\ldots (2)$$

where n denotes the sum of generators and a_i, b_i, c_i are the fuel coefficients cost of the ith unit

We must also consider the valve opens constraints, because it has the ability to fuel cost due to wire drawing effect. It can easily change the objective function to have non-differentiable points. Inorder to handle this issues a sinusoidal function is added and the equation is

$$F_t = \sum_{i=1}^{n} F_i(p_{gi}) = \sum_{i=1}^{n}\sum_{i=1}^{n} [a_i (p_{gi})^2 + b_i (p_{gi}) + c_i + e_i * |\sin(f_i * (p_{gi}^{min} - p_{gi}))|] \ldots\ldots (3)$$

Generation capacity constraints

The above mentioned objective function is subjected to the below displayed constraints.

Operation limit unit

The main constraints of power generation of each power unit should not be exceed its higher and lower limits

$$p_{gi}^{LL} \le p_{gi} \le p_{gi}^{UL} \ldots\ldots (4)$$

where p_{gi} is the amount of output power unit for the ith generator

p_{gi}^{LL} *is the lower limit of power output of ith generator*

p_{gi}^{UL} *is the upper limit of power output of ith generator*

Power Balance constraints

The total produced powers consider be equal to the total losses and total demand load. The mathematical expression for power balance is

$$\sum_{i=1}^{n} p_{gi} = p_D + p_L \ldots\ldots (5)$$

p_L *refers to* total line loss and p_D refers to the system demand. The Kron's formula for finding the transmission loss is refers to the system demand. The Kron's formula for finding the transmission loss is

$$p_L = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{gi} B_{ij} p_{gi} + \sum_{i=1}^{n} B_{0i} p_{gi} + B_{00} \ldots\ldots (6)$$

B_{ij} *denotes the* $(i-j)$*th element of loss coeffcient of symetric matrix B*

B_{0i} *ith element of loss coefficient vector of the symetric matrix B*

B_{00} *is the loss coeffcient for normal function conditions*

ARTIFICIAL BEE COLONY OPTIMIZATION

For millions of years, Social insects are living on earth. The nest is constructed; production of solution is organized and fetches food[13]. The social insects in the colonies are highly flexible, it can easily fix to the environment which is alters occasionally. Hence, social insects of the colony need to be strong and it should have the ability to handle disturbance occurred disturbances [14]. The interactions are calculated multitude of variety of signal generated by chemical and/or physical signals. The final outcome of various actions and interactions [15] denotes the behavior of social insect colony. The dance is staged to obtain the food and the performance of dance denotes the direction of the food source and it the signal is generated to other insects in colony [16]. The bees have the ability to travel long distance in order to collect the food sources. In general, flower patches with good amounts of nectar is gathered by consuming less effort , where other nectar is less can be receive fewer bees [17]. The looting process initiates by scout bees and when they are sent to find for effective flower patches and it moves from one flower to another. The scout bee initiated the waggle dance in front of bees to describe the better food sources to provide better communication [18]. Once the process of dance is completed the dancer will return back to flower patch with follower bees. The scout bee which has

more number of followers are consider to be more proficient patches, where it can easily collects the food. In ABC algorithm, the population of bees classified into two groups, scout and employed bees. The scout bees look out for a new food source and the role employed bees is to identify a food source within the neighborhood. The sharing of information with other bees is carried in a better way. The [Fig. 1] shows the flowchart of Artificial Bee Colony Optimization algorithm with ns scout bees randomly distributed in the search space. The quality of nectar amounts of sites is considered to be the visited sites of scout bees. The sites which contain high amount of nectar are selected for neighborhood search. To have proper exploration the recruit bee is selected and the nectar amounts are calculated with high nectar amounts. Select m sites which have the highest nectar amounts from sites to form the next bee population. The onlooker bees calculate the information about the food sources based on the probability value using

$$P_i = \frac{fit_i}{\sum_{j=1}^{N_{FS}} fit_i} \ldots\ldots\ldots\ldots (7)$$

where fit_i denotes fitness value of the solution i , and N_{FS} is the sum of food sources that are same to the employed bees quantity, ne. Now the onlookers generate a modification in the chosen position and evaluate the nectar amount of the new source:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \ldots\ldots\ldots\ldots (8)$$

K$\in$ {1, 2, 3,...., N_{FS}} and j$\in$ {1, 2, 3,...., D} are randomly chosen. The value of K should be different from i and it need to be random. ϕ_{ij} is random and it is selected from $[-1,1]$. The onlooker bee updates the new position, when the quantity of nectar of the new source is greater than that of the previous nectar amount. If this condition fails, then that food source is falls under abandoned, pairing of the employed bee with that particular food source becomes a scout. The abandoned source can be calculated by using the equation

$$x_{ij} = x_{jmin} + rand\,(0,1) * \phi_{ij}(x_{jmax} - x_{jmin}) \ldots\ldots\ldots\ldots (9)$$

x_{jmin} x_{jmax} *are the lower and upper bound* Limit of the parameter to be optimized. The main control parameters in the ABC algorithm are the sun of employed bees, onlooker bees, the Limit Value, and the maximum cycle number [12]. In ABC algorithm the process of balancing the exploration and exploitation process is done by employed, onlooker bees and scouts bees.

Pseudo code of EABC algorithm for EDP

In this section, an ABC algorithm is used to resolve the optimal output of each generating power unit for a specified demand, in order to reduce the power loss and to reduce total generation cost [19]. The possible solution is described by the position of a food source. The amount nectar of a food source depends to the quality or fitness of the associated solution [20]. The procedure for the proposed method is as described as.

Step1: Determine the generator cost efficient, generated power limits and parameters of ABC are initialized.

Step2: The population is initialized with random generated solutions and the dimension D is mapped number of generator unit G , the size of the population is described as N. Each solution consist of Xi=[Pi1,Pi2,...,Pij...PiD] where i=(1,2,....,N) and j=(1,2,...,D), here D is referred to number of generators of the population N. The each solution of elements is referred to xij is the real power output of generating units (Pgi).

$$P_{ij} = P_{j\,min} + rand(0,1) * (P_{j\,max} - P_{j\,min}) \ldots\ldots\ldots\ldots (10)$$

Step3: The fitness calculation for the population is generated for each position of food equivalent to employed bees in the colony. Fix the cycle count as one and the following steps repeated until the MCN is done.

$$fit_i = \frac{1}{1+f_i} \text{ if } f_i \geq 0 \ldots\ldots\ldots\ldots (11)$$

$$fit_i = 1 + abs(f_i) \text{if } f_i < 0 \ldots\ldots\ldots\ldots (12)$$

Step 4: The employed Bees illustrates about the changes of position and site selection for identifying the new food source. Inorder to find the new food source to improve the exploration is

$$v_{i,j} = (x_{h1,j} + x_{h2,j})/2 + \phi_{i,j}(x_{h1,j} - x_{h2,j}) * \mathrm{M} + \varphi_{i,j}(x_{best,j} - x_{h1,j}) \text{ where}$$

$$\mathrm{M} = \frac{f(x_{h2}) - f(x_{h1})}{f(x_{h1}) + f(x_{h2})} \ldots\ldots\ldots\ldots (13)$$

where $h1$ *and* $h2$ are various integers chosen in [1,SN]. $\phi_{i,j}$ *and* $\varphi_{i,j}$ *is a random number gnereated unfirmoly within* $[0,0.5]$ *and* $[0,1]$, $x_{best,j}$ to find the best solution for jth element. Here midpoints between the two random points are calculated. By using the above equation exploitation is improved and the performance of the exploration is enhanced. The two random

IIOAB JOURNAL ISSN: 0976-3104

integer's parameters $\phi_{i,j}$ and $\varphi_{i,j}$ which controls the step size for search. The exploration process is carried if the value of $\phi_{i,j}$ *is higher* and if the exploitation is carried if the value of $\phi_{i,j}$ is smaller.The range [0, 0.5] is defined inorder to avoid the search in the local optima. $\varphi_{i,j}$ is a random number generated uniformly within [0,1] to makes the search process contributes to the current best solution. The comparison of old position with fitness value for the changed position is computed

Step 5: It impose the modifications of position by onlookers, here step 4 procedure gives input to the onlookers bee. The amount of nectar for the candidate source is calculated and it checks old nectar source, if the new one is better than previous one, then it is restored. If not the previous one is retained and greedy selection mechanism is incorporated for carrying the selection process.

Step 6: An Abandon source exploited by the bees and this process is carried if a solution is not improvised in a predefined number of trials. The scout bee process is initiated to identify new food source to be replaced with previous one. This process is carried by using the equation

$$x_{ij} = x_{jmin} + rand\,(0,1) * \phi_{ij}\left(x_{jmax} - x_{jmin}\right) \ldots \ldots \ldots \ldots \ldots (14)$$

Step 7: The best solution is memorized and the cycle count is incremented.

Step 8: Once the termination constraints are achieved, ending of the process is initiated if the termination criteria are satisfied. If the process is not completed go to Step 4 to find best fitness and food source with equivalent position. Then the termination criteria is choose as the optimum output powers of generating units for ED procedure for that time interval.

RESULTS AND DISCUSSION

The ABC algorithm is incorporated for practical applications and it is tested on different test cases such as 6, 15, 20 and 38 unit systems. The [Table 1] represent the parameters settings used for all systems. Here we consider transmission cost, transmission loss for all four cases. A moderate B-loss coefficients matrix of power system network is used, for evaluating the transmission loss and computational time. A system with core i7 and 8 GB of RAM is utilized to run the code in Matlab 16.

Table 1: Parameters setting used with Test Systems

S.No	Number of Iterations	Population Size	C
6 -unit	1200	10	0.2
15-Unit	3000	10	0.2
38-Unit	13000	10	0.2

Case 1: Six Unit System

In this system consist of six power generating units with the total load demand on the system. The constrained power system load demand (*PD*) is mentioned 1263 MW and the coefficients of fuel cost [14,18].The loss coefficients in the power transmission line (matrix *B*) are shown. In this case, randomly generated of P1,P2,...,P6 of generator power output for each individual is calculated. The population is mapped and it is compared with other variants of SA [21], TS, PSO[22], GA, and NPSO-LRSto the proposed EABC. The best results are compared with the above mentioned algorithm and the result is shown in the [Table 2]. It shows the proposed method satisfy the constraints of system constraints, such as generator power limits and prohibited zones of units [23]. [Table 2] listed the statistic results that involved the minimum fuel cost and transmission loss generation cost, evaluation value.

Table 2: Simulation Results for 6 Unit Systems

Method	Min. (S/hr)	Mean (S]hr)	Max. (S/hr)	TFE	Time (sec.)	Standard deviation
SA [15]	15461.1	15488.98	15545.5	NA	50.36	28.367
TS[15]	15454.89	15472.56	15498.05	NA	20.55	13.719
PSO[15]	15450.14	15465.83	15491.71	1,00,000	6.82	10.15
GA[16]	15459	15469	15524	20,000	41.58	0.057
NPSO-LRS[17]	15450	15454	15492	20,000	14.89	0.002
EABC	15443.2	15450.3	15499.4	10,000	4.32	0.001

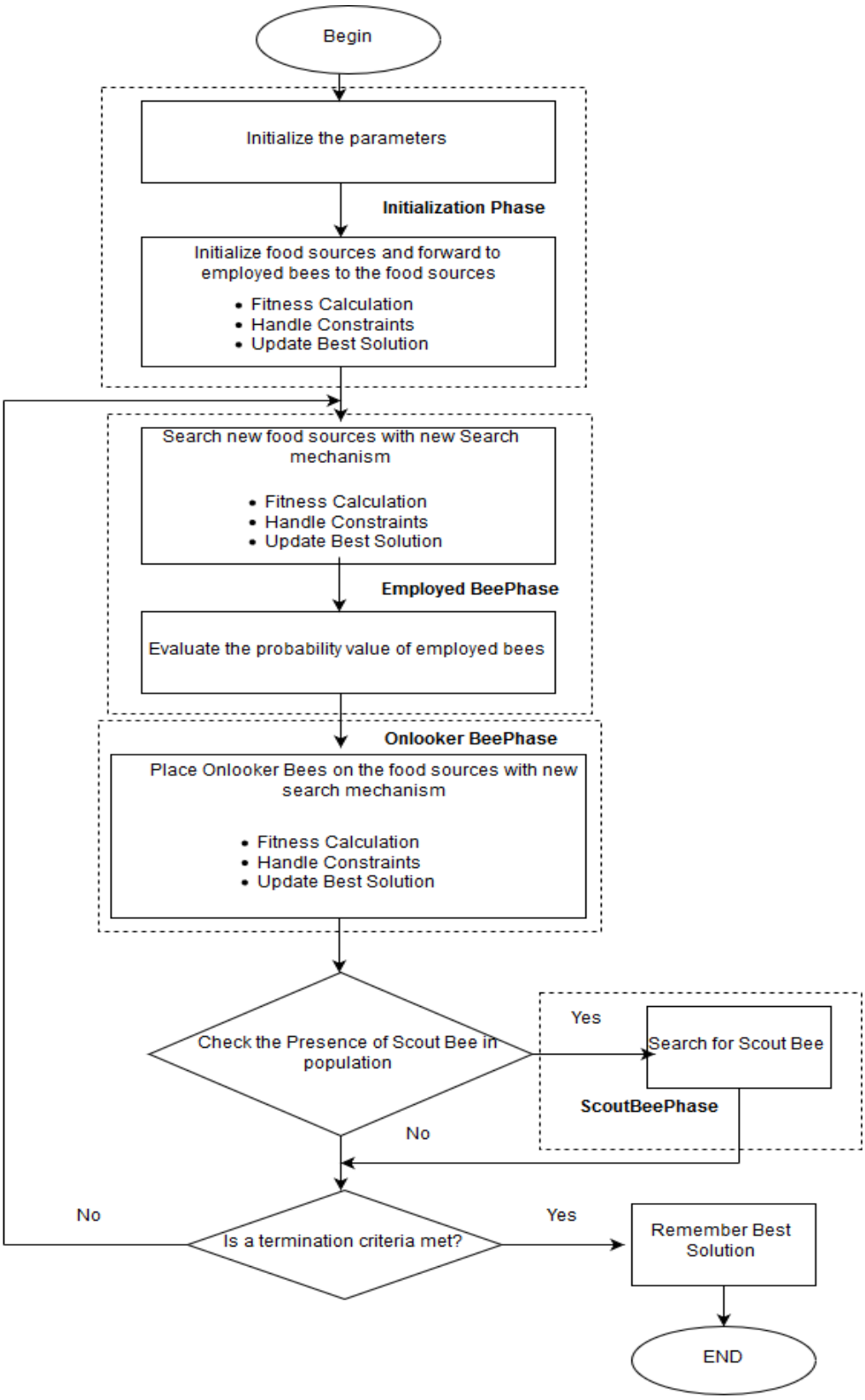


Fig. 1: Workflow for enhanced ABCO

IIOAB JOURNAL
ISSN: 0976-3104

Case 2: 20 Unit systems

In Case 2 the 20 generating system is incorporated with the constraints such as generating load demand and the transmission loss. The system input data is referred from the valve loading effect is not taken to the count, but the loss of transmission is noted. The total demand load for this use case is 2500 MW. The best results are compared with the algorithm and the outcome is shown in the [Table 3]. It shows the proposed method satisfy the constraints of system constraints, such as generator power limits and prohibited zones of units. [Table 3] listed the statistic results that involved the minimum fuel cost and transmission loss generation cost, evaluation value [23]. It is compared to the other Evolutionary algorithms [24,25] for effective results.

Table 3: Simulation Results for 20 Unit Systems (2500 MW)

Unit Power output(MW)	BBO	LI	HM	IABC	GABC	EABC
P1	513.09	512.78	512.78	325.3754	277.4268	360.2834
P2	173.35	169.10	169.10	174.2311	87.40052	113.9078
P3	126.92	126.89	126.89	50	50	50
P4	103.33	102.87	102.87	50	65.6027	50
P5	113.77	113.64	113.68	78.18726	121.0221	81.74645
P6	73.07	73.57	73.57	20	28.50667	28.57458
P7	114.98	115.29	115.29	118.1033	118.0878	118.0681
P8	116.42	116.40	116.40	50	50	50
P9	100.69	100.41	100.41	121.1041	156.9833	93.05378
P10	100.00	106.03	106.03	43.54856	77.50294	30
P11	148.98	150.24	150.24	241.1593	241.1436	241.0568
P12	294.02	292.76	292.76	423.9738	422.4694	425.9739
P13	119.58	119.12	119.12	96.81424	150.1453	152.0103
P14	30.55	30.83	30.83	78.48479	53.49366	99.72673
P15	116.45	115.81	115.81	168.2717	81.21983	117.4553
P16	36.23	36.25	36.25	33.32164	33.32311	33.33886
P17	66.86	66.86	66.86	49.89571	39.7736	35.56004
P18	88.55	87.97	87.97	30	50.86036	44.00757
P19	100.98	100.80	100.80	43.0362	43.02828	105.8806
P20	54.27	54.31	54.31	30	30	30
Total generation (MW)	2592.10	2591.97	2591.97	2225.507	2177.99	2260.644
Total transmission loss (MW)	92.10	91.97	91.97	274.4929	322.01	239.356
Total generation cost (S/h)	62456.78	62456.64	62456.63	60213.47	60216.32	60191.41

Case 3:38 Unit systems

The 38 generating system is used without valve point loading and other two constraints such as generating load demand and the transmission loss. The input data is referred from the [26] and for the load demand is 6000 for MW. The best results are compared with the algorithm and the outcome is shown in the [Table 4]. It shows the proposed method satisfy the constraints of system constraints, such as generator power limits and prohibited zones of units [27,28]. [Table 3] listed the statistic results that involved the minimum fuel cost and transmission loss generation cost, evaluation value.

Table 4: Simulation Results for 38 Unit Systems (6000MW)

Unit Power output(MW)	DE/BBO	BBO	PSO-TVAC	NEW-PSO	EP-EPSO	EABC
P1	426.60606	422.230586	443.659	550	318.0777	349.5643
P2	426.606054	422.117933	342.956	512.263	475.117	324.0509
P3	429.663164	435.779411	433.117	485.733	399.1265	325.8652
P4	429.663181	445.48195	500	391.083	500	500
P5	429.663193	428.475752	410.539	443.846	500	327.0012
P6	429.663164	428.649254	492.864	358.398	500	327.3221
P7	429.663185	428.119288	409.483	415.729	500	326.9571
P8	429.663168	429.900663	446.079	320.816	500	326.9238
P9	114	115.904947	119.566	115.347	114	114
P10	114	114.115368	137.274	204.422	132.7826	114

P11	119.768	115.418662	138.933	114	114	114
P12	127.0728	127.511404	155.401	249.197	114	114
P13	110	110.000948	121.719	118.886	110	110
P14	90	90.0217671	90.924	102.802	90	90
P15	82	82	97.941	89.039	82	82
P16	120	120.038496	128.106	120	120	120
P17	159.598	160.303835	189.108	156.562	141.9435	147.2366
P18	65	65.0001141	65	84.265	65	65
P19	65	65.000137	65	65.041	65	65
P20	272	271.999591	267.422	151.104	120	272
P21	272	271.872268	221.383	226.344	272	272
P22	260	259.732054	130.804	209.298	260	260
P23	130.648618	125.993076	124.269	85.719	80	96.42769
P24	10	10.4134771	11.535	10	10	10
P25	113.305034	109.417723	77.103	60	92.9577	85.23205
P26	88.0669159	89.3772664	55.018	90.489	55	72.21293
P27	37.5051018	36.4110655	75	39.67	35	35
P28	20	20.009888	21.682	20	20	21.20601
P29	20	20.0089554	29.829	20.985	20	20
P30	20	20	20.326	22.81	20	20
P31	20	20	20	20	20	20
P32	20	20.0033959	21.84	20.416	20	20
P33	25	25.0066586	25.62	25	25	25
P34	18	18.0222107	24.261	21.319	18	18
P35	8	8.0000426	9.667	9.122	8	8
P36	25	25.006066	25	25.184	25	25
P37	21.782	22.0005641	31.642	20	38	20
P38	21.0621792	20.6076309	29.935	25.104	20	20
Cost ($/h)	94,17,235.79 [21]	94,17,633.64 [21]	95,00,448.31 [22]	95,96,448.31 [22]	93,87,925.50 [22]	90,13,940

CONCLUSION

In this paper, an Enhanced Artificial Bee colony optimization is discussed and this searching technique provides better convergence, where generally the optimization algorithms should have the ability to balance both exploration and exploitation. The algorithm helps to improve the exploration and it avoids trapping in the local optimum. This method is incorporated for solving the complex problem in Economic dispatch. Here we impose the proposed algorithm in three cases inorder to prove its performance such as 6 generation unit system, 15 generation unit system and 38 generation power system. The proposed method helps to minimize the transmission loss and reduce the generation cost with short computational time. The results are compared with other traditional approaches and the displayed results illustrates that the proposed algorithm outperforms the results of other approaches. It tells the proposed algorithm can be easily used to solve the complex issues occur in the economic dispatch.

CONFLICT OF INTEREST
There is no conflict of interest

ACKNOWLEDGEMENTS
This work is a part of the Research Projects sponsored by Visvesvaraya Ph.D Scheme for Electronics & IT, Ministry of Electronics & Information Technology, India, and ReferenceNos: PHD-MLA-4(44)/2015-16, dated August 2015. The authors would like to express their thanks for the financial supports offered by the Sponsored Agency.

FINANCIAL DISCLOSURE
No financial support was received for this study.

REFERENCES

[1] Lu, Peng, et al.[2014] Chaotic differential bee colony optimization algorithm for dynamic economic dispatch problem with valve-point effects. International Journal of Electrical Power & Energy Systems 62: 130-143.

[2] Gaing Zwe-Lee.[2003] Particle swarm optimization to solving the economic dispatch considering the generator constraints. IEEE transactions on power systems 18(3): 1187-1195.

[3] Zhou Jianzhong, et al. [2017] A multi-objective multi-population ant colony optimization for economic emission dispatch considering power system security." Applied Mathematical Modelling.

[4] Basu M. [(2013] Artificial bee colony optimization for multi-area economic dispatch. International Journal of Electrical Power & Energy Systems 49: 181-187.
[5] Jadhav HT, Ranjit Roy. [2013] Gbest guided artificial bee colony algorithm for environmental/economic dispatch considering wind power. Expert Systems with Applications 40(16): 6385-6399.
[6] Sen Tanuj, Hitesh Datt Mathur.[2016] A new approach to solve Economic Dispatch problem using a Hybrid ACO–ABC–HS optimization algorithm. International Journal of Electrical Power & Energy Systems 78: 735-744.
[7] Abdelaziz AY, Ali ES, Abd Elazim SM. [2016] Flower pollination algorithm to solve combined economic and emission dispatch problems. Engineering Science and Technology, an International Journal 19(2): 980-990.
[8] Secui, Dinu Calin. [2015] The chaotic global best artificial bee colony algorithm for the multi-area economic/emission dispatch. Energy 93: 2518-2545.
[9] Adriane BS. [2013] Cuckoo search for solving economic dispatch load problem. Intelligent Control and Automation
[10] Parouha, Raghav Prasad, Kedar Nath Das.[2016] DPD: An intelligent parallel hybrid algorithm for economic load dispatch problems with various practical constraints. Expert Systems with Applications 63: 295-309.
[11] James JQ, Victor OK Li.[2016] A social spider algorithm for solving the non-convex economic load dispatch problem." Neurocomputing 171: 955-965.
[12] Song, Xiaoyu, Qifeng Yan, Ming Zhao.[2017] An adaptive artificial bee colony algorithm based on objective function value information." Applied Soft Computing 55: 384-401.
[13] Hemamalini S, Sishaj P. Simon.[2010] Artificial bee colony algorithm for economic load dispatch problem with non-smooth cost functions." Electric Power Components and Systems 38(7): 786-803.
[14] Elsayed WT, et al.[2016] Modified social spider algorithm for solving the economic dispatch problem. Engineering Science and Technology, an International Journal 19(4): 1672-1681.
[15] Saravuth Pothiya, Issarachai Ngamroo, Waree Kongprawechnon,[2008] Application of multiple tabu search algorithm to solve dynamic economic dispatch considering generator constraints, Energy Convers. Manage. 49: 506–516.
[16] Gaing ZL.[2003] Particle swarm optimization to solving the economic dispatch considering the generator constraints, IEEE Trans. Power Syst. 18 (3) :1187–1195.
[17] Selvakumar K. Thanushkodi. [2007] A new particle swarm optimization solution to nonconvex economic dispatch problems, IEEE Trans. Power Syst. 22 (1) :42–51
[18] Kaushal, Rajanish Kumar, Tilak Thakur, and Isarar Ahamad. [2016] SOLUTION OF ECONOMIC LOAD DISPATCH PROBLEM INCLUDING LINE LOSSES USING BAT ALGORITHM. 8(1):
[19] Shaw, Binod, et al. [2011] Solution of economic load dispatch problems by a novel seeker optimization algorithm." International Journal on Electrical Engineering and Informatics 3(1): 26.
[20] Bhattacharya PK. Chattopadhyay R,[2010] Hybrid differential evolution with biogeography-based optimization for solution of economic load dispatch, IEEE Transactions on Power Systems, 25: 1955–1964
[21] Vlachogiannis JK, Lee KY.[2009] Economic load dispatch – A comparative study on heuristic optimization techniques with an improved coordinated aggregation-based PSO, IEEE Transactions on Power Systems, 24: 991– 1001
[22] PANDIAN, SEVUGARATHINAM MUTHU VIJAYA, and Keppanagowder Thanushkodi. Considering transmission loss for an economic dispatch problem without valve-point loading using an EP-EPSO algorithm. TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES 20.Sup. 2 (2012): 1259-1267.
[23] Behera R, Pati BB, Panigrahi BP. [2011] Economic power dispatch problem using artificial immune system." International Journal of Scientific &Engineering Research 2.5
[24] Pandi V. Ravikumar, et al.[2011] Economic load dispatch using hybrid swarm intelligence based harmony search algorithm. Electric power components and systems 39(8): 751-767.
[25] Swain RK, Sahu NC, Hota PK.[2012] Gravitational search algorithm for optimal economic dispatch. Procedia Technology 6: 411-419.
[26] Anusha B, Noah, Sivaranjani C, Priyanka S,[2015] Predictive analysis of movie reviews using hybrid approach", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 – 8907, 1(1): 1-7
[27] Govindharaj I, Karthiga S, Manishalakshmi R, Mary Silvia Theodore R.[2016] Home Power Analyzer with Smart Power Monitoring using IoT, International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907,2(1): 7-13
[28] Rajadurai R, Amelia, Aubrey, A.Anusha, Danapriya P, Geethashnee D. [2017] Efficient Data Leakage Prevention Strategy using Key Distribution", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907,3(1): 9-16.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

SOLVING ECONOMIC POWER DISPATCH PROBLEM WITH TRANSMISSION LOSS AND VALVE POINTS USING WHALE OPTIMIZATION ALGORITHM

Rajeswari Muniyan[1]*, J. Amudhavel[2], Dhavachelvan Ponnurangam[1]
[1]*Department of Computer Science, Pondicherry University, Puducherry, INDIA*
[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT

Economic Power Dispatch (EPD) is a non-linear and non-convex optimization problem in the field of power system for operational planning. EPD is the sharing of power demand among generator units by minimizing cost while satisfying constraints. This work presents an investigation on meta-heuristic approach Whale Optimization algorithm (WOA) to solve EPD problem. To demonstrate the efficacy of the proposed algorithm two benchmark test problems having ten-unit system and thirteen-unit system are evaluated. The simulation results are compared with other nature inspired algorithms and Whale optimization algorithm proves its superior performance in terms of convergence rate and accuracy.

KEY WORDS
Evolutionary algorithm, economic power dispatch problem, power system, Whale optimization algorithm, Transmission Loss.

Received: 23 June 2017
Accepted: 20 July 2017
Published: 15 Sept 2017

***Corresponding Author**
Email:
raji.rajeswari18@gmail.com

INTRODUCTION

ED problem is allocating power to the generating units for minimized cost with constraints. The generators are coordinated so that lowest cost generator is used as much as possible rather than costlier generator. Costlier generators are used when the demand is increased. Traditional ED problem are solved using conventional methods such as Lamba iteration method, gradient method, Lagrangian multiplier method, base point participation factors method and branch and bound method. In these numerical methods, incremental fuel cost curves of generating units are required which in turn monotonically increase piece-wise linear cost function approximation. The input-output characteristic of non-linear ED includes ramp-rate, valve point, prohibited zones and fuel cost functions which are non-convex in nature. Thus it get trapped in multiple local minimum points and for larger-scale generating units conventional method results in longer computational time due to oscillatory problem. Dynamic programming [1] is used to solve non-linear economic dispatch problem with valve point. However this method had dimensional problem for large-scale generating units and results in local optimality.

In past decades, the researcher found alternative solution to conventional methods for solving ED problem is meta-heuristic optimization algorithms like evolutionary programming [2], genetic algorithm [3], simulated annealing [4], tabu search [5], particle swarm optimization [6]. This method does not require large memory and their iterative search strategy helps to find optimal solution by eliminating local optima. The selection of control parameters for evolutionary algorithms helps to attain optimal solution. The other stochastic search algorithms uses artificial intelligent techniques like Hopfield neural network [7], Adaptive Hopfield neural network [8]. These methods have sigmoidal functions which results in huge numerical calculations and iterations.

Hybrid algorithm started to emerge due to the quality of the solution. Hybridization of conventional PSO with generator constraints is used to solve ED problem. The diversification process in PSO was improved with the use of differential evolution based mutation operator. This algorithm takes the advantage of both method and provides robust result [9]. Dipankar Santra et al [10] proposed hybrid PSO-ACO algorithm to solve Economic Load Dispatch (ELD) problem by considering transmission loss. They provided solution for both convex and non-convex ELD problems using 3-generator 5-bus system. They considered ramp rate, valve value, prohibited zone, transmission loss and capacity of the generator. Their hybrid approach handles smooth and non-smooth, convex and non-convex problems. Boubakeur Hadji et al [11] proposed a variant of dance bee colony algorithm with dynamic step size adjustment to solve economic emission dispatch problem with valve effects. The performance of the proposed variant is tested on IEEE 30-Bus and 40 unit generators.

D.B. Prakash, C. Lakshminarayana [12] proposes a Whale Optimization Algorithm to increase the reliability and stability of the power system. The proposed approach reduce the operational cost with inequality constraints. SeyedaliMirjalili and Andrew Lewis [13] proposed nature-inspired meta-heuristic optimization algorithm called Whale Optimization Algorithm (WOA). It mimics the natural behavior of hunting of preys in humpback whales. The technique used to carry out the hunting process is called bubble-net. It is shown that the algorithm is compared with other meta-heuristic algorithms as well as conventional algorithms.

The remainder of this work is organized as follows: Section 2 presents the formulation of the economic power dispatch problem. Section 3 provides the overview and algorithm framework to solve EPD. The simulation results of a ten-unit problem and a thirteen-unit problem are shown in Section 4. Finally Section 5 concludes the work with summary.

IIOAB JOURNAL
ISSN: 0976-3104

ECONOMIC POWER DISPATCH PROBLEM

Economic dispatch problem may be of static or dynamic. The static economic dispatch problem allows constant power supply for the given time with minimized generation cost. In the static method, the generation load of the power system is to be determined for the scheduled period and it may vary based on the customer demand. The nature of the static ED problem the load should be scheduled according to the customer demand. The dynamic economic dispatch, dynamic constraints are used to retain the life of the generator with the help of ramp rate units to minimize the cost of generation unit. The ramp rate unit is to maintain the life of the power generator by avoiding shortening of the generator and thermal stress. Solving dynamic ED problem is diffiult due to large dimensional but provides the accurate formulation.
The objective function of the dynamic ED problem is to minimize the cost of generation unit of the output power system. The generation cost (C_G) of the power system is the summation of the cost of each generating unit N and the mathematical formulation is given as follows.

$$Minimize\ C_G = \sum_{i=1}^{N} C_i(P_i) \quad (1)$$

where N is the number of generators, C_i and P_i is the cost function and power of the ith generator in the power systems.
The fuel cost of generating unit is represented in the quadratic polynomial function of output power as follows.

$$C_i(P_i) = x_i P_i^2 + y_i P_i + z_i \quad (2)$$

where x_i, y_i and z_i are the cost coefficient of the ith generator.
The quadratic polynomial function represented is smooth function in practical which cannot determine the input or output of the power generators. Value point effects consists of higher-order nonlinearity sinusoidal function are added with generation cost function to obtain an accurate result.

$$C_i(P_i) = x_i P_i^2 + y_i P_i + z_i + V_i \quad (3)$$

where V_i is the higher-order nonlinearity sinusoidal function caused by value point effects and it can be defined as

$$V_i = \left|u_i \sin(v_i(P_i^{min} - P_i))\right| \quad (4)$$

where u_i, v_i are value point coefficients of generator i and P_i^{min} is the minimum generation limit.
The objective function of the ED problem is represented as

$$Minimize\ C_G = \sum_{i=1}^{N} x_i P_i^2 + y_i P_i + z_i + V_i + \left|u_i \sin(v_i(P_i^{min} - P_i))\right| \quad (5)$$

Constraints
The equality and inequality constraints to be satisfied in economic dispatch problem in terms of generating capacity and power balance are described.
Power balance
The total generated power is the summation of the total power demand and transmission loss in the power system. The constraint is mathematically formulated as.

$$\sum_{i=1}^{N} P_i = P_D + P_{TL} \quad (6)$$

where P_D is the total power demand in the system, P_{TL} is the total transmission loss in the system.
The total transmission loss is calculated using B coefficient and the formulated is represented as follows.

$$P_{TL} = \sum_{i=1}^{N}\sum_{j=1}^{N} P_i B_{ij} P_j + \sum_{i=1}^{N} B_{0i} P_i + B_{00} \quad (7)$$

where B_{ij}, B_{00} and B_{0i} are the ijth loss coefficient of the matrix B, loss coefficient constant and ith element of loss coefficient vector respectively.
Generating capacity constraint
The power output of the each generation must satisfy generating capacity constraints and it is represented as.

$$P_i^{min} \le P \le P_i^{max} \quad (8)$$

where P_i^{min} and P_i^{max} are the minimum limit and maximum limit of the power system output of generator i.

PROPOSED SYSTEM

Whale optimization algorithm mimics the social behavior of bubble-net hunting in humpback whales. This algorithm describes the special hunting behavior called bubble-net feeding method. During hunting the whales follow 9-shaped path or creation of circulation movement. During hunting, humpback whale moves down to the water at 15meter depth. It knows the position of the prey and produce bubbles in spiral shape which encircles the prey. Whale hunting can be defined in three different ways namely, Encircling the prey, Bubble-net feeding and search for a prey.

IIOAB JOURNAL
ISSN: 0976-3104

Encircling prey

Humpback whales can identify the exact location of the prey i.e small fishes and it update its current position towards optimal solution for each iteration. Thus for unknown result in search space, the current best solution is prey for each iteration. Once the optimal solution is obtained it is updated in position table and it can be used by other search agents. To update the position according to the current best solution can be found using

$$\vec{D} = |\vec{C}.\vec{X}^*(t) - \vec{X}(t)| \quad (9)$$

where $\vec{D}$ represents distance vector between current position ($\vec{X}$) and best position $\vec{X}^*$ for each iteration t. $\vec{C}$ represents the coefficient vector and provides direction towards current best solution. The solution for next iteration can be found using

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A}.\vec{D} \quad (10)$$

where $\vec{A}$ represents coefficient vector. The coefficient vectors are calculated using

$$\vec{A} = 2\vec{a}.\vec{r} - \vec{a} \quad (11)$$

$$\vec{C} = 2.\vec{r} \quad (12)$$

where $\vec{a}$ is the linearly decreased in range [2,0] with respect to increase in iteration and $\vec{r}$ is random vector range [0,1].

Bubble-Net Feeding

Bubble-net behavior of whale can be in two ways as follows, shrinking encircling and spiral updating position.

Shrinking Encircling

The value of a is decreased from 2 to 0 for the iteration. In this method, whales does not exhibit discontinued circle motion between its current and predecessor position. It exhibits continuous spiral path to hunt prey.

Spiral updating position

The spiral updating position between whale and prey forms helix-shape movement as

$$\vec{X}(t+1) = \vec{D}'.e^{bl}.\cos(2\pi l) + \vec{X}^*(t) \quad (13)$$

where $\vec{D} = |\vec{X}^*(t) - \vec{X}(t)|$ shows the distance of whale to prey, b is defined as constant value to define the shape of the logarithmic spiral and l is the random number range from [-1,1].

SeyedaliMirjalili, et al. [18] shows the probability of 50-50% which follows either shrinking encircling or spiral model using

$$\vec{X}(t+1) = \begin{cases} \vec{X}^*(t) - \vec{A}.\vec{D} & if\ p < 0.5 \\ \vec{D}'.e^{bl}.\cos(2\pi l) + \vec{X}^*(t) & if\ P \geq 0.5 \end{cases} \quad (14)$$

where p is the random number [0,1].

Search for prey

Humpback whales to search for prey randomly according to the position of each other search agents. The exploration of search space to find best solution for each iteration can be calculated using

$$\vec{D} = |\vec{C}.\overrightarrow{X_{rand}} - \vec{X}| \quad (15)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A}.\vec{D} \quad (16)$$

EXPERIMENTAL DESIGN AND ANALYSIS

Experimental environment setup

In order to prove the performance of the proposed algorithm on solving EPD problem, two different benchmark test systems: ten-unit system [14] and thirteen-unit system [15] are considered. The experiments are conducted on different nature inspired algorithms under similar environment conditions

IIOAB JOURNAL
ISSN: 0976-3104

in order to evaluate the performance of WOA. The proposed algorithm is coded using MATLAB 2015 platform is used to solve small and medium scale EDP problem by WOA under Windows on an Intel 2 GHz Core 2 quad processor with 2GB RAM. [Table 1] describes the instances considered by WOA to solve EDP. The empirical evaluations will set the parameters of the proposed system. Appropriate parameter values are determined based on the preliminary experiments.

Table 1: Parameter settings for Experimental Evaluation

Type	Method
Number of Bees	100
Maximum Iterations	1000
Heuristic Used	Modified Nelder-Mead Method
Termination Condition	Maximum Iterations
Run	20
$\vec{a}$	Range =(2,0)
l	Range =(-1,1)

Case Study

Test case 1

In this case study the benchmark [25- 27] system of ten-unit with total demand of 300MW is tested. Along with power balance and generating capacity constraints valve point effect and transmission loss are considered. The cost coefficient and B-loss coefficient matrix are given in [14]. The efficiency of the WOA to solve test case 1 are compared with the other optimization algorithm like ACO [16], ABC [17], GSA [18], and HAS [19] along with traditional approach Gradient search and PSO [6]. From the [Table 2], it is evident that our proposed algorithm WOA outperforms with respect to fuel cost.

Table 2: Unit output of different methods for test case 1

Unit	WOA	ABC	ABC	GSA	HSA	PSO	Gradient search
P1	49.77446	27.74896	94.33846	97.52896	36.48311	0	0
P2	13	13	13	13.00106	13	0	0
P3	10.00713	10	10	10	10	20.4046	18.56
P4	15	15.03227	15.09696	15	15.00042	15	15
P5	21.56978	21.48645	21.43165	21.41054	21.3972	65.2755	125
P6	14	14.00178	14.00092	14.00387	14.03365	44.0719	34.9027
P7	20	20.00383	20.04302	20	20	20	68.7361
P8	25	25	25	25.02442	25	29.7851	0
P9	150	150	150	150	150	17.5	0
P10	20.00144	20.05393	20.00574	20.13838	20.14682	96.1525	44.9138
Total Loss	38.35282	16.32722	82.91674	86.10724	25.06118	0.849413	0.730886
Total Cost	7367.335	7367.794	7367.854	7368.862	7368.607	12343.774	11274.296

Test case 2

The system comprises of thirteen-unit benchmark test case with valve units and transmission loss is taken as test case 2. The overall load demand of the test case is 2520MW. The data for cost coefficient and B-loss coefficient matrix have been extracted from [14]. The performance of the WOA to solve test case 2 are compared with variants of Grey wolf OGWO [20], GWO [21] and other nature inspired meta-heuristic algorithms OIWO [22], SDE [23], ORCCRO [24]. On comparing with existing algorithms, our proposed WOA outperforms with respect to fuel cost are shown in [Table 3].

Table 3: Unit output of different methods for test case 2

Unit	WOA	OGWO	GWO	OIWO	SDE	ORCCRO
P1	454.5374	628.2940	628.1678	628.3185	628.32	628.32
P2	344.3787	299.1803	298.9229	299.1989	299.20	299.20
P3	358.8063	297.5041	298.2269	299.1991	299.20	299.20
P4	179.2704	159.7284	159.7232	159.7331	159.73	159.73
P5	173.2179	159.7325	159.7210	159.7331	159.73	159.73
P6	174.4962	159.7295	159.7270	159.7331	159.73	159.73
P7	178.5798	159.7334	159.7173	159.7330	159.73	159.73
P8	145.1762	159.7323	159.6793	159.7331	159.73	159.73
P9	143.607	159.7327	159.6673	159.7330	77.40	77.40
P10	96.9857	77.3963	77.3971	77.3953	113.12	112.14
P11	107.2662	114.7487	114.6051	113.1079	92.40	92.40
P12	74.82939	92.3974	92.3886	92.3594	92.40	92.40
P13	91	92.3780	92.3550	92.3911	92.40	92.40
Total Loss	22.151	40.2874	40.2983	40.3686	40.43	39.43
Total Cost	24324.03	24512.7250	24514.4774	24514.83	24514.88	24513.91

CONCLUSION

This work solves economic power dispatch problem for power units by considering transmission loss and valve points using WOA. WOA is implemented to solve small and medium sized benchmark unit systems such as ten-unit system and thirteen unit system. To solve EPD, the social behavior of humpback whales and the hunting of prey by the whales are mapped to obtain optimal solution. To evaluate the performance of the WOA, the obtained results are compared with other existing algorithms and the performance are tabulated in Table 2 and 3. From the simulation result, it is evident WOA outperforms in succeeding minimal cost for the power generation unit for various test cases.

CONFLICT OF INTEREST

There is no conflict of interest

ACKNOWLEDGEMENTS

This work is a part of the Research Projects sponsored by the Major Project Scheme, UGC, India, Reference Nos: F.No./2014-15/NFO-2014-15-OBC-PON-3843/ (SA-III/WEBSITE), dated March 2015. The authors would like to express their thanks for the financial supports offered by the Sponsored Agency.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Liang Z-X, Duncan Glover J. [1992] A zoom feature for a dynamic programming solution to economic dispatch including transmission losses. IEEE Transactions on Power Systems 7(2): 544-550.

[2] Sinha Nidul, Chakrabarti R, Chattopadhyay PK. [2003] Evolutionary programming techniques for economic load dispatch. IEEE Transactions on evolutionary computation 7 (1): 83-94.

[3] Chiang C-L.[2007] Genetic-based algorithm for power economic load dispatch. IET generation, transmission & distribution 1(2): 261-269.

[4] Wong, Kit Po, Yin Wa Wong.[1994] Genetic and genetic/simulated-annealing approaches to economic dispatch. IEE Proceedings-Generation, Transmission and Distribution 141(5): 507-513.

[5] Lin, Whei-Min, Fu-Sheng Cheng, and Ming-Tong Tsay. [2002] An improved tabu search for economic dispatch with multiple minima. IEEE Transactions on power systems 17(1): 108-112.

[6] Park Jong-Bae, Ki-Song Lee, Joong-Rin Shin, Kwang Y. Lee. [2005] A particle swarm optimization for economic dispatch with nonsmooth cost functions." IEEE Transactions on Power systems 20(1): 34-42.

[7] Park JH, Kim YS, Eom IK, Lee KY. [1993] Economic load dispatch for piecewise quadratic cost function using Hopfield neural network. IEEE Transactions on Power Systems 8(3): 1030-1038.

[8] Lee, Kwang Y, Arthit Sode-Yome, June Ho Park. [1998] Adaptive Hopfield neural networks for economic load dispatch,IEEE Transactions on Power Systems 13(2):519-526.

[9] Khamsawang S, Wannakarn P, Jiriwibhakorn S. [2010] Hybrid PSO-DE for solving the economic dispatch problem with generator constraints." In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, 5:135-139. IEEE.

[10] Santra, Dipankar, Anirban Mukherjee, Krishna Sarker, and Debasish Chatterjee. [2016] Hybrid PSO-ACO algorithm to

solve economic load dispatch problem with transmission loss for small scale power system. In Intelligent Control Power and Instrumentation (ICICPI), International Conference on, pp. 226-230. IEEE
[11] Hadji, Boubakeur, Belkacem Mahdad, Kamel Srairi, Nabil Mancer. [2015] Multi-objective Economic Emission Dispatch Solution Using Dance Bee Colony with Dynamic Step Size. Energy Procedia 74: 65-76.
[12] Prakash DB, Lakshminarayana C. [2016]Optimal siting of capacitors in radial distribution network using Whale Optimization Algorithm. Alexandria Engineering Journal
[13] Mirjalili, Seyedali, and Andrew Lewis. [2016]The whale optimization algorithm. Advances in Engineering Software 95: 51-67.
[14] Sen Tanuj, Hitesh Datt Mathur. [2016] A new approach to solve Economic Dispatch problem using a Hybrid ACO–ABC–HS optimization algorithm. International Journal of Electrical Power & Energy Systems 78: 735-744.
[15] Črepinšek Matej, Shih-Hsi Liu, Marjan Mernik. [2013] Exploration and exploitation in evolutionary algorithms: A survey. ACM Computing Surveys (CSUR), 45(3)
[16] Mohd Rozely Kalil, Muhammad Khayat Idris, Titik Khawa Abdul Rahman, Mohd Rafi Adzman.[2009] Ant colony optimization (aco) technique in economic power dispatch problems." In Trends in Communication Technologies and Engineering Science, pp. 191-203. Springer Netherlands,
[17] Labbi, Yacine, Djilani Ben Attous, Belkacem Mahdad. [2014]Artificial bee colony optimization for economic dispatch with valve point effect. Frontiers in Energy 8(4): 449-458.
[18] Duman Serhat, Aysen Basa Arsoy, Nuran Yörükeren. [2011] Solution of economic dispatch problem using gravitational search algorithm. In Electrical and Electronics Engineering (ELECO), 2011 7th International Conference on, pp. I-54. IEEE.
[19] Pandi VR, Panigrahi BK, Mallick MK, Abraham A, Das S. [2009] August. Improved harmony search for economic power dispatch. In Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on 3:-408. IEEE.
[20] Pradhan, Moumita, Provas Kumar Roy, Tandra Pal. [2017] Oppositional based grey wolf optimization algorithm for economic dispatch problem of power system. Ain Shams Engineering Journal.
[21] Wong, Lo Ing, Sulaiman MH, Mohamed MR, Mee Song Hong. [2014] Grey Wolf Optimizer for solving economic dispatch problems." In Power and Energy (PECon), 2014 IEEE International Conference on, pp. 150-154. IEEE
[22] Bhattacharjee, Kuntal, Aniruddha Bhattacharya, and Sunita Halder nee Dey. "Oppositional real coded chemical reaction optimization for different economic dispatch problems." International Journal of Electrical Power & Energy Systems 55 (2014): 378-391.
[23] Ciornei, Irina, and Elias Kyriakides.[2012] A GA-API solution for the economic dispatch of generation in power system operation. IEEE Transactions on power systems 27(1): 233-242.
[24] Bhattacharya, Aniruddha, Pranab Kumar Chattopadhyay. [2010] Biogeography-based optimization for different economic load dispatch problems. IEEE transactions on power systems 25(2): 1064-1077.
[25] Anusha B, Noah C. Sivaranjani, Priyanka S.[2015] "Predictive analysis of movie reviews using hybrid approach", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 1(1):1-7.
[26] Govindharaj I, Karthiga S, Manishalakshmi R, Mary Silvia Theodore R.[2016]Home Power Analyzer with Smart Power Monitoring using IoT", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 2(1): 7-13, 2016.
[27] Rajadurai R, Amelia Aubrey, Anusha A, Danapriya P, Geethashnee D. [2017] Efficient Data Leakage Prevention Strategy using Key Distribution", International Research Journal of Advanced Engineering Sciences and Technologies, ISSN: 2455 - 8907, 3(1): 9-16.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A SURVEY ON MULTI-OBJECTIVE TRAVELLING SALSMAN PROBLEM

Kanimozhi Jayamoorthi[1*], Dinesh Karunanidy[2], Amudhavel Jayavel[2], Subramanian Ramalingam[1]

[1]*Department of Computer Science and Engineering, Pondicherry University, Puducherry, INDIA*
[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT

Traveling Travelling Salesman Problem (TSP) is a challenging problem in combinatorial optimization. The important of this problem is due to the fact that it is used in many fields such as transportation, logistics, semiconductor industry, problem of routing, scan chain optimization and drilling problem in integrated orbit test, production and many others scientific and industrial fields. In this paper we consider the multi-objective TSP (MOTSP) which is an extended instance of traveling salesman problem (TSP). Since TSP is a NP- hard problem MOTSP is also NP-hard problem. In MOTSP simultaneous optimization of more than one objective functions is required. TSP considers and optimizes one objective function to find the best solution. Instead in MOTSP many objectives are considered and optimized to find the best solutions. Many algorithms are used to solve MOTSP. Some algorithms give optimal solution and some give the nearest optimal solution. By evolving a population of solutions, multi-objective evolutionary algorithms (MOEAs) are able to approximate the Pareto optimal set in a single run. It results in nearest optimal solution within a reasonable amount of time by optimizing many objectives. In this survey, we review the current state-of-the-art computational algorithms used to solve the MOTSP.

KEY WORDS
Multi-objective travelling salesman problem, Multi-objective evolutionary algorithm, NSGA II, Decomposition, Ant colony optimization

INTRODUCTION

The Travelling Salesman Problem (TSP) is an optimization problem used to find the shortest path to travel through the given number of cities. Travelling salesman problem states that given a number of cities N and the distance or time to travel between the cities, the traveler has to travel through all the given cities exactly once and return to the same city from where he started and also the cost of the path is minimized. This path is called as the tour and the path length or travel time is the cost of the path [1] [2]. The TSP mathematical model follows [3] [2]:

$$\min Z = \sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}C_{ij}$$

Subject to $\sum_{i=1}^{N} X_{ij} = 1$ j=1,2,3.....N

$\sum_{j=1}^{N} X_{ij} = 1$ j=1,2,3..... N

Let C_{ij} be the cost for traveling from i-th city to j-th city. Where Xij = 1 if the salesman travels from city-i to city-j, otherwise Xij = 0. In the mathematical model it is shown that the distance is considered for finding the best solution.

The travelling salesman problem can be classified as symmetric Travelling Salesman Problem (STSP), Asymmetric Travelling Salesman Problem (ATSP), and Multi Travelling Salesman Problem (MTSP) [9] [1]. In travelling salesman problem with increasing the number of cities the existing solutions don't provide optimal solution at the appropriate time. Moreover the solution found by optimizing the distance to travel to the destination may not always give the best solutions. For the solution optimization more than one objective needs to be considered. When more objectives functions are considered to optimize the solutions then it will give the better solutions compared to solutions given by single objective function [5]. This gives the need for MOTSP. MOTSP considers more than one objective functions to optimize the solutions. In MOTSP, the aim is to simultaneously optimize several conflicting objectives, such as shortest travelling distance, minimum time, minimum cost, and lowest risk. Under multi-objective framework no single point is considered as an optimal solution, because an improvement in one objective will cause at least another objective not being able to be optimized. Therefore, the optimal solution can only be a set of non-dominated and trade-off solutions. The [6] gives the mathematical model of MOTSP by considering two objectives for optimization.

The MOTSP can be formulated as a multi-objective model with two objective functions. The first objective function (1) considers the minimization of the distance traveled by the salesman, while the second objective function (2) considers the working time of the salesman.

$$(1) \quad \min Z1 = \sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}C_{ij}$$

Received: 3 June 2017
Accepted: 20 July 2017
Published: 15 Sept 2017

Corresponding Author
Email:
janathakani@gmail.com
Tel.: +91 7373231828

IIOAB JOURNAL
ISSN: 0976-3104

$$(2) \quad \min Z2 = \sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}T_{ij}$$

Subject to $\sum_{i=1}^{N} X_{ij} = 1$ j=1,2,3.....N

$\sum_{j=1}^{N} X_{ij} = 1$ j=1,2,3..... N

Let C_{ij} be the cost for traveling from i-th city to j-th city. T_{ij} is the travel cost from i-th city to j-th city and X_{ij} = 1 if the salesman travels from city-i to city-j, otherwise X_{ij} = 0. Both the objectives are optimized to find the best solutions.

MOTSP can be solved using conventional technique and Evolutionary based technique. [Fig. 1] classifies the techniques to solve MOTSP. Under conventional technique there are two ways to solve it. 1) Weighted Sum Technique: This technique converts multiple objectives into single objective using linear combination of objectives. But it requires a prior knowledge of weightage of each objective of a problem. 2) Constraint Based Technique: This technique considers only one objective at a time and treats remaining k-1 objectives as constraints. Final answer is computed by taking average of results obtained for all objectives. Application of this technique demands a prior knowledge of constraints of the problem. Due to these reasons both are not the best techniques to solve the MOTSP.

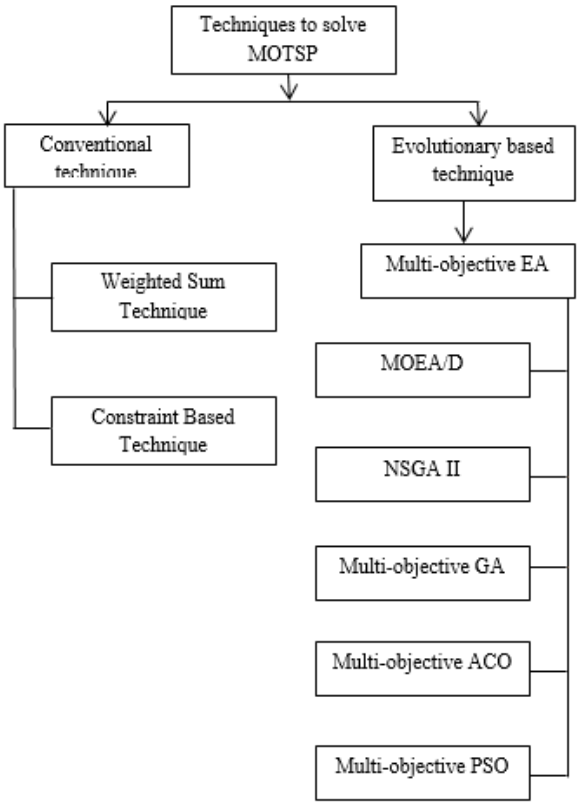


Fig. 1: Techniques to solve MOTSP

Under evolutionary based technique Multi-objective evolutionary algorithms (MOEAs) are well-suited for solving several complex multi-objective problems with more objectives. MOEA generates a set of non-dominated solutions at the end of each run. MOEA generates a set of non-dominated solutions at the end of each run, which is called Pareto set. The Pareto front contains set of Pareto solutions. Generally, an external archive is used by MOEAs to maintain a set of non-dominated Pareto set solutions [7] [4]. MOEA can also be used to solve single objective problems with or without constrained [67]. In addition to that bi-objective problems are solved using MOEA in [9] [10] [11] [12]. In [13] given the reasons to use MOEAs for solving multi-objective optimization problems :(i) they are easy to implement, (ii) MOEAs return more than one optimal solution, (iii) there are less chances of the algorithm to get stuck in local minima, (iv) MOEAs are flexible and robust, (v) MOEAs do not require any a prior knowledge of the problem.

As in [13] MOEAs can be classified into elitist and non-elitist algorithms. Elitist MOEAs have a mechanism to preserve good solutions at every generation while non-elitist MOEAs do not have such mechanism. Therefore, non-elitist algorithms perform worst monotonically than elitist algorithms. In MOEAs, a solution x is called non-dominated solution if it is better in all objectives than the solution y or solution x is strictly

IIOAB JOURNAL
ISSN: 0976-3104

better than the solution y in at least one objective. The solutions which do not satisfy above two conditions are called dominated solutions. In MOEAs, at the end of each generation, we have a set of non-dominated solutions. When we plot these non-dominated solutions on a graph, such graph is called Pareto front and Solutions on the Pareto front are called Pareto set (solutions). At the termination of MOEA, we have a set of non-dominated solutions. We obtained optimal Pareto front at the end of termination of MOEA and solutions on it are called optimal Pareto solutions.

MATERIALS AND METHODS

In this section we present a short overview of state-of-the-art computational algorithms as well as the working of those algorithms. Under Multiobjective Evolutionary Algorithm (EA) many algorithms are used to solve MOTSP. The following factors make the difference in solving the problem. 1) Expected fitness of the solution 2) The uncertainty of the solutions 3) Algorithm convergence speed 4) Usage of external archive. All these above factors give the different ways of solving the MOTSP. The categories of MOEA given by the above factors are follows:

State of art Computational Algorithms:

a. Multi-objective Evolutionary Algorithm based on Decomposition (MOEA/D)
b. Multi-objective Genetic Algorithm
c. NSGA II
d. Multi-objective Ant colony optimization
e. Multi-objective Particle swarm optimization

Multi-objective Evolutionary Algorithm based on Decomposition (MOEA/D)

MOEA/D [16] is a general EA framework for dealing with MOPs. Like generic MOEAs, a MOEA/D starts with an initial population of candidate solutions, in an iteration it generates some new trial solutions and selects the fittest ones to the next iteration; and it repeats the process until some termination conditions are satisfied. One of the major advantages of MOEA/D is that it is very easy to design local search operator within it using well-developed single-objective optimization algorithms. In MOEA/D, an MOTSP is decomposed into a set of scalar objective sub-problems and a probabilistic model, using both priori and learned information, is built to guide the search for each sub-problem. By the cooperation of neighbor sub-problems, MOEA/D could optimize all the sub-problems simultaneously and thus find an approximation to the original MOTSP in a single run [30] [31] [16]. This idea is realized by solution cooperation in neighborhood, i.e., the solutions in the same neighborhood are used to generate new trial solutions, and the new trial solutions only update the old solutions in the same neighborhood. The following two notations are extremely important in an MOEA/D [17].

Sub-problem: a multi-objective optimization problem is decomposed into a set of scalar objective problems and each of them is called a sub-problem. Hopefully, the optimal solution of the i^{th} sub-problem $g^i(\pi)$ lies in the PS (PF) of the original problem.

Neighborhood: The neighborhood $B_i = (i_1, i_2, \ldots, i_K)$ of the i^{th} sub-problem contains the indices of similar sub-problems, i.e., the i_j^{th} ($j = 1, \ldots, K$) sub-problems are the most similar ones to the i^{th} sub-problem.

Sub-problem definition: In [16] Tchebycheff approach is used to define the sub-problems as given follows:

$$\min g^i(\pi) = g\left(\pi|\lambda^i, z^*\right) = \max_{1<j\leq m} \lambda_j^i \left| f_i(\pi) - z_j^* \right|$$

Where $\lambda_i = (\lambda^i_1, \ldots, \lambda^i_m)^T$ is a weight vector with the i^{th} sub-problem, $z* = (z*1, \ldots, z*m)^T$ is a reference point. Reference point weakly dominates all the other solutions in the population. It is clear that all the sub-problems are with the same form and can only be differentiated by the weight vectors. If two vectors are close to each other, the corresponding sub-problems should be similar to each other and their optima should also be close in both the decision and objective spaces in most cases. By using the weight vectors, the neighborhood could be determined before the algorithm execution. A Probabilistic model Pi stores information extracted from the population for the i^{th} sub-problem and mating process will be continued. MOEA/D framework follows.

MOEA/D Algorithm framework

The following steps are followed in the MOEA/D algorithm framework [18] [16]. First MOEA/D converts an MOTSP into N sub-problems and randomly generates a solution for each sub-problem. Then initialize the reference point $z*$ followed by Initialize the weight vectors, neighborhood and probability matrix for each sub-problem. For each sub-problem i = 1,..., N, do Sample a new solution be a unvisited city by randomly selecting according to the probability and repeat the process until the whole tour is constructed. Then do update of reference point, Update solutions Set and Update probability model. If the stopping conditions are satisfied, then stop; otherwise do the process again. Algorithm 1 explains the MOEA/D working process.

IIOAB JOURNAL
ISSN: 0976-3104

Algorithm 1: MOEA/D

Input: N: number of SOSPs (scalar optimization sub-problems) ; W: number of the neighbors for each SOSP; $\lambda^1, \ldots, \lambda^N$: uniformly distributed weight vectors; pc: crossover rate; pm: mutation rate.

Initialization
1. Set EP = Ø (External Population)
2. For each _i, calculate the W closest weight vectors, $\lambda^{i(1)}, \ldots, \lambda^{i(W)}$, by Euclidean distance and set φ(i) = {i(1), . . ., i(W)}.
3. Generate an initial population $x_1, \ldots x_N$ and evaluate fu(xj)for each individual
4. Initialize z = (z1,..., zm)
Repeat
5. For i = 1 to N do
Reproduction:
6. Generate a new solution y by two individuals x_u and x_l using crossover and mutation operators, where u, l ∈ φ(i)
Improvement:
7. Improve y by using a problem-specific improvement repair operator, which is optional.
Update of z:
8. For j = 1,..., m, if fj(y) < zj, set zj= fj(y)
9. Update of neighboring solutions
Update of EP:
10. Remove those solutions dominated by y from EP and add y to EP if it is not dominated by any member in EP
Termination:
11. Until stopping criteria are satisfied, output EP

In [19], Algorithm decomposes the population into 's' scalar optimization sub-problems according to the Tchebycheff approach. It follows a two-chromosome representation for individual's representation. The first chromosome locates the cities while the second chromosome indicates which vehicle is to be assigned to visit the city specified in the first chromosome, which improves the performance of the algorithm. A chemical reaction optimization based decomposition method is introduced in [6]. It follows the same steps as MOEA/D algorithm framework; in addition to that a chemical collision stage is followed after initialization stage. A Parallel Procedure for Dynamic Multi-objective TSP [21] follows the decomposition method to decompose the problems and follow the parallel execution of the sub-problems for time efficient process. In [21] objectives are represented in the form of matrix, in which we can increase the number of objectives for optimization. But the number of objectives affects directly the execution time. In [31] different approaches in MOEA/D are listed to solve the MOTSP.

Multi-objective Genetic Algorithm

Multi Objective Genetic Algorithm (MOGA) was given by Fonseca et al [4] [13]. It is an extension of single objective optimization algorithm. The rank to an individual is assigned based on the number of solutions in the population by which it is dominated. GA has two main parts, an evolution function and a fitness function. In the case of the MOTSP, the parameters produced by the evolution function might be the order of the nodes through which the path will go. The fitness function in that same case would return the total length of the path found. The GA would then compare fitness values for each input string and assign priority to the ones that returns lower path lengths and other objectives. Based on the objective values the best solutions will be produced. The framework of the MOEA was explained in [14] [24]. By following it MOGA steps are explained below.

Framework of MOGA

MOGA starts with initial population and assign the fitness value to the individuals and apply the Evolution operator on the individuals. Do this process until satisfaction criteria met and finally it will produce the best path for MOTSP. Algorithm 2 explains the MOGA working process.

Algorithm 2: MOGA

Input: $Population_{size}$, $Problem_{size}$, $P_{crossover}$, $P_{mutation}$
Output: S_{best}
Population <- InitializePopulation ($Population_{size}$, $Problem_{size}$)
EvaluatePopulation (Population)
S_{best} <- GetBestSolution (Population)
While (~StopCondition())
    Parents <- SelectParents (Population, $Population_{size}$)

```
        Children <- Ø
        For (Parent1, Parent2 € Parents)
                Child1, Child2 <- Crossover (Parent1, Parent2, Pcrossover)
                Children <- Mutate (Child1, Pmutation)
Children <- Mutate (Child2, Pmutation)
        End
        EvaluatePopulation (Children)
        Sbest <- GetBestSolution (Children)
        Population <- Replace (Population, Children)
End
Return (Sbest)
```

GA can solve problems with non-parametrical problems, multi-dimensional, non-continuous and non-differential optimization problems. But Genetic Algorithm has weaker local search ability [14]. During the later period of Hierarchical Genetic Algorithm, the fitness converges, and less superior individuals are produced. However, Hybrid Simulated Annealing Algorithm can make it jump out of the erroneous zone of local optimum. Due to the compatibility of Genetic Algorithm, it is feasible to combine Simulated Annealing Algorithm and Hierarchical Genetic Algorithm to form the modified Simulated Annealing Genetic Algorithm [8]. The modified Simulated Annealing Genetic Algorithm can sooner achieve a better global optimum solution. The reasons are: the suffix structure design of chromosome reduced the space of the solution, the self-adaptive genetic operator and double crossover and mutation improved 'premature convergence problem'; the introduction of Simulated Annealing Algorithm stretched the fitness and enhanced the local search ability.

Multiple traveling salesman problem was solved using GA in [5]. Two types of GAs were developed in it. The first one is a multi-objective GA that uses the sum of the route distance of each salesman to estimate the overall distance, and the standard deviation of the routes to estimate the balance. The second is a mono-objective GA with a fitness function that combines these two objectives by the use of a parameter. GA is modified into discrete GA and combined with fuzzy technique [34] in order to solve multi-objective problem. The proposed algorithm is able to find better result in shorter computational time. Xiaomei Sun [69] solved the Intelligent Transportation Systems (ITS) using improved Genetic Algorithm. It divides the population into subgroups and do the selection operation and recombine the subgroups followed by crossover and mutation operation. In [19] MOGA is used to solve vehicle routing problem.

Non-dominated Sorting Genetic Algorithm-II (NSGAII)

Non-dominated Sorting Genetic Algorithm-II (NSGAII) was introduced by Deb et al. [13]. It is an improved version of NSGA. The rank of every solution is computed based on how many number of solutions it dominates. In order to maintain the diversity of a population the algorithm finds average distance of two neighbors on either side of a solution along each of the objectives. The calculated distance is called crowding distance of that solution. For generating mating pool for next generation, selection of solutions is performed based on rank and crowding distance. The concept of dominance is used to find the best individual. It follows:

Concept of dominance

The concept of dominance is applied to multi-objective problems to compare two solution candidates $X1$, $X2$, and determine if a solution dominates the other one. In particular, the dominance is a method for the classification of the solutions which ensures the selection of the best solution in the resulting population Rt.

Definition [6]: Given two solutions $X1$ and $X2$, solution $X1$ dominates solution $X2$, if the following conditions are satisfied:
1. Solution $X1$ is not worse than $X2$ for all the objectives;
2. Solution $X1$ is strictly better than $X2$ for at least one objective.

Using these dominance concept individuals are ranked. When two solutions have the same rank then a solution that has higher crowding distance is selected for mating. The algorithm selects the solutions for the next generation based on following policy: select best solutions out of union of the best parents and best offspring (obtained after application of genetic operators). The following criteria are used for selection of the best solutions from the union: fitness and spread. As the algorithm selects the best solutions from the union, it does not require extra memory to preserve elite solutions. The NSGA II algorithm framework follows [27] [28] [29].

NSGA II Algorithm Framework

It starts with the initial population and assigns the rank for each individual. Then crowding distance is calculated for each individual. Based on the rank and crowding distance of the individuals best solutions are selected for crossover and mutation operations. Until the stopping criteria met the steps are repeated.

IIOAB JOURNAL
ISSN: 0976-3104

Algorithm 3: NSGA II

Input: Population $_{size}$, problemSize, P $_{crossover}$, P $_{mutation}$
Output: children

```
Population <- InitializePopulation (Population_Size , ProblemSize)
EvaluateAgainstObjectiveFunctions(Population)
FastNondominatedSort(Poplation)
Selected <- SelectParentsByRank(Population, Population size )
Children <- CrossoveAndMutation (Selected, P crossover, P mutation)
While (~ StopCondition())
        EvaluateAgainstObjectiveFunctions(Children)
        Union <- Merge (Population, Children)
        Fronts <- FastNondominatedSort (Union)
        Parents <- Ø
Front L<- Ø

For (Front i € Fronts)
CrowdingDistanceAssignment (Front i)
If (Size (Parents) + Size (Fronts i) > Population size)
Front_L <- i
Break()
      Else
Parents <- Merge (Parents, Front i)
      End
End
If (Size (Parents) + Size (Fronts i) < Population size)
Front L <- SortByRankAndDistance (Front L)
        For (P1 to P_Population size – SizeFront L)
    Parents <- Pi
End
End
Selected <- SelectPrentByRankAndDistance (Parents, Population Size)
Population <- Children
Children <- CrossoverAndMutation(Selected, P_crossover, P_mutation
End
Return (Children)
```

In [6], Author proposed a methodology based on the non-dominated sorting genetic algorithm NSGA-II to solve the Multiple TSP. It follows the same algorithm framework given above but the number of salesman will be more than one. An improved NSGA II is implemented in [40]. Specifically, a layer strategy according to need is proposed to avoid generating unnecessary non-dominated fronts. The arena's principle is also adopted to construct non-dominated set, so as to reduce the count of dominance comparison. In addition, an order crossover like operator and an inversion mutation operator are adopted for MOTSP.

Non-dominated sorting differential evolution algorithm for the minimization of route based fuel consumption multi-objective vehicle routing problems is introduced in [31] [29]. It uses the hybrid version of NSGA II to solve the routing problem by optimizing the fuel consumption. MOTSP is solved using two algorithms based on Differential Evolution, and the third one is based on NSGA II. The algorithms solve 2 to 5 objective functions TSP in [27] [28]. Vehicle routing problem with uncertain travel cost is solved using NSGA II in [5] [48].

Multi-objective Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) is a population-based meta-heuristic inspired by the pheromone trail laying and following behavior of some real ant species. ACO was originally designed for solving single-objective combinatorial optimization, and later it has proven to be one of the most successful algorithms for modeling discrete optimization problems. Due to notable results on these applications, ACO algorithms were soon extended to tackle problems with more than one objective (MOPs), called multi-objective ACO (MOACO), most of these algorithms have different design choices and focus in terms of Pareto optimally, that is, they do not make a priori assumptions about the decisions maker's preferences. So it gives the best result in real time problem solving.

ACO implementation mainly consists of two stages: tour construction and pheromone update. The process of tour construction and pheromone update is applied iteratively until a termination condition is met (such as a set number of iterations). Both the tour construction and pheromone update stages can be performed independently [17]. The main idea in this algorithm is that the behavior of each individual ant produces an emergent behavior in the colony. When applied to the MOTSP, individual agents ("ants") traverse the graph of the problem, leaving a chemical (pheromone) trail behind them. At each node it comes to, an ant

must decide which edge to take to the next node. This is done by checking each edge for pheromone concentration and applying a probability function to the decision of which edge to choose. The higher the concentration of pheromone, the more likely the ant is to choose that edge. Also, to avoid stagnation in travel, the pheromone is given an evaporation rate so that at each iteration the pheromone loses a certain percentage on each edge. The MOACO algorithm framework is explains the above given steps to solve MOTSP.

MOACO Algorithm Framework

The algorithm maintains: (I) $\tau^1, \ldots, \tau^K$, where τ^j is the current pheromone matrix for group j, storing its learned knowledge about the sub-region of PF that it aims at approximating; (II) $\eta 1, \ldots, \eta^N$, where η^i is the heuristic information matrix for sub-problem i, which is predetermined before the construction solution starts; (III) EP, which is the external archive containing all the non-dominated solutions found so far. Algorithm 4 explains the working process of MOACO.

Algorithm 4: MOACO

```
Define Original pheromone
        For iteration=1 to i= (1,2,3...n)
          For ant=1 to i = (1,2,3...n)
            Random start node
                Do while node < i
                    Select next node
                Loop
            Next End Loop ant
            Calculate Multi – objective function
    For ant=1 to i=(1,2,3...n)
                Do while node < i
                   Update pheromone
                Loop
                Checking pheromone upper and pheromone lower
          Next End Loop ant
        Next End loop iteration
```

All the problems using MOACO follow the above MOACO algorithm framework. MOACO can be combined with Genetic Algorithm and MOEA/D to give better performance. In [42], a hybridized algorithmic approach to solve 4- dimensional Travelling Salesman Problem is introduced. The algorithm is a hybridization of rough set based ant colony optimization (rACO) with genetic algorithm (GA). In this the initial solutions are produced by ACO which act as a selection operation of GA and then GA is developed with a virgin extended rough set based selection (7-point scale), comparison crossover and generation dependent mutation. The development of rACO-GA is in general form and it can be applied in other discrete problems such as network optimization, graph theory, solid transportation problems, vehicle routing, covering salesman problem, VLSI chip design, industrial information integration, information management, supply chain, airline industry, etc. ACO is combined with PSO and used to solve TSP which gives good result [2].

In [30] [24] ACO is implemented to solve multi objective problems combined with Genetic Algorithm. Following other MOEA/D-like algorithms, MOEA/D-ACO decomposes a multi-objective optimization problem into a number of single-objective optimization problems. Each ant (i.e., agent) is responsible for solving one sub-problem. All the ants are divided into a few groups, and each ant has several neighboring ants. An ant group maintains a pheromone matrix, and an individual ant has a heuristic information matrix. During the search, each ant also records the best solution found so far for its sub-problem. To construct a new solution, an ant combines information from its group's pheromone matrix, its own heuristic information matrix, and its current solution. An ant checks the new solutions constructed by itself and its neighbors, and updates its current solution if it has found a better one in terms of its own objective. ACO is used to solve Multi-objective Bus Route Planning.

Multi-objective Particle Swarm Optimization (MOPSO)

Particle Swarm Optimization (PSO) was given by (Kennedy and Eberhart 1995) [15]. This method is inspired by the social behavior of animals living swarm. Initially, they tried to simulate the ability of birds to fly synchronously and their ability to change direction suddenly while remaining optimal training. The particles are individuals and they move into hyperspace research based on limited information:

1. Each particle has a memory which enables to store the best point at which it has already passed and it tends to return to that point.
2. Each particle is informed of the best known point in its neighborhood and it will tend to go to that point.

The movement of a particle tends to follow her go her own way or particle tends to return to the best site in which it has already passed or the Particle tends to move toward the best ever achieved by its neighbors. The position of each particle is modified according to its own experience and that of its neighbors. To make the PSO capable to handle MOP the following modifications are done.

IIOAB JOURNAL
ISSN: 0976-3104

General modifications on PSO to handle MOPs are: (1) External Archive Maintenance (2) Select Global Leaders (3) Update Personal Best (4) Mutation Operator (Perturbation). PSO produce a single solution and the solution of a problem with multiple objectives is not a single solution (as in global optimization). Instead, in the multi-objective optimization, we aim to find different solutions (called Pareto Optimal Set) using MOPSO. By following these steps MOPSO framework is explained below.

MOPSO framework

Initialize swarm population and velocity. Do the Fitness evaluation and Pareto dominance for ranking particles (solutions). Based on the ranking of individuals Personal Best (Pbest) swarm population is stored on to the memory and non-dominated solutions are stored in External Archive. Particles Select global leaders from External Archive and compute PSO equation. Using fitness rank individuals are mutated. The new Personal Best are updated and maintained on external archive. If satisfaction criteria met then stop the process else continue [15]. Algorithm 5 explains the working of MOPSO.

Algorithm 5: MOPSO

```
Begin
Initialize Swarm
Initialize leaders in an external Archive
Quality (Leaders)
g=0;
While g < gmax
    For each Particle
            Select leader
            Update Position
            Mutation
Evaluation
Update pbest
   End For
Update leaders in the external Archive
Quality (leaders)
g++
End while
Report results in the external Archive
End
```

In [15] MOTSP is solved using MOPSO. In multi-objective Particle Swarm Optimization (MOPSO), a ranking operator is applied to the population in a predefined iteration to build an initial archive using €-dominance to select the best solutions for mating. It uses the advantage of Pareto approach which is based on the concept of external archive with the rapidity of convergence of PSO as to minimize the total distance traveled by a particle and minimize the total time.

RESULTS

This section studies the effectiveness of the listed five MOEA algorithms. Then, we evaluate the overall performance of MOEA algorithms by comparing its Performance measures. The associated parameters are defined below.

Performance measures

To evaluate the performance of the MOEA algorithms, we employ the following widely recognized performance measuring metrics. Let PF_{ref} be a reference set of solutions well approximating the true PF, and PF_{known} be the set of non-dominated solutions obtained by an algorithm.

Approximation set [34]: An approximation set is defined by Zitzler et al. as follows: let A $\in H$ be a set of objective vectors. A is called an approximation set if any element of A does not dominate or is not equal to any other objective vector in A. The set of all approximation sets is denoted as Z. The result of solving a real-world problem usually is an approximation set A and not the Pareto optimal front PF*.
Cardinality metrics: the cardinality of A refers to the number of solutions that exists in A. Intuitively; a larger number of solutions are preferred.
Accuracy metrics: this aspect refers directly to the convergence of A. In other words, it indicates how distant A is from the theoretical Pareto optimal front PF*. Notice that when the Pareto optimal front is unknown, a reference set R is considered instead.
Diversity metrics: distribution and spread are two very closely related facets. The distribution refers to the relative distance among solutions in A, while the spread refers to the range of values covered by the solutions in A. The spread is also known as "the extent" of an approximation set.

IIOAB JOURNAL
ISSN: 0976-3104

Maximum spread (MS) [34]: MS reflects how well the true PF is covered by the points in PF_{known} through the hyperboxes formed by the extreme function values observed in PF_{ref} and PF_{known}.
Hypervolume (HV): HV also known as S metric, hyper-area is an unary metric that measures the size of the objective space covered by an approximation set. A reference point must be used to calculate the mentioned covered space. HV considers all three aspects: accuracy, diversity and cardinality.
Inverted generational distance (IGD) [34]: IGD is defined as the average distance from each point v in PF_{ref} to its nearest counter-part in PF_{known}, as follows:

$$IGD = \frac{\sum_{v \in PF_{ref}} d(v, PF_{known})}{|PF_{ref}|}$$

Where $d(v, PF_{known})$ is the Euclidean distance (in the objective space)between solution v in PF_{ref} and its nearest solution in PF_{known} and $|PF_{ref}|$ is the number of solutions in PF_{ref}. IGD measures the convergence and diversity of an obtained non-dominated solution set. This metric is commonly used and a lower IGD indicates a better overall performance of an algorithm.
Average Computational Time (ACT): ACT is the average running time consumed by an algorithm over 30 runs. This metric is a direct indicator of the computational complexity of an algorithm.

Result comparison

IGD reflects the overall performance of an algorithm regarding the quality of the obtained PF_{known}. So we compared all the listed MOEA algorithms using IGD. We computed the IGD of the algorithms for 10 runs with randomly generated cities. The IGD of MOEA/D, NSGA II, MOGA, MOACO and MOPSO algorithms are shown in the [Table 1]. RGC represents Randomly Generated Cities.

Table 1: Results of IGD of the MOEA algorithms

RGC/ Algorithms	MOEA/D	NSGA II	MOGA	MOACO	MOPSO
1	0.00	0.00	0.89	1.50	2.07
2	0.00	0.29	1.33	1.72	2.39
3	0.00	0.50	1.42	2.06	2.46
4	0.54	1.08	1.68	2.38	2.48
5	0.77	1.22	2.01	2.26	2.56
6	0.80	1.50	2.10	2.53	2.63
7	0.76	1.36	2.34	2.56	2.96
8	0.62	1.25	2.56	3.06	3.76
9	0.17	1.46	2.60	3.16	3.96
10	0.20	1.37	2.50	3.28	4.45

As mentioned in the IGD definition, lower IGD indicates a better overall performance of an algorithm. From the [Table 1], it is clear that MOEA/D is having the lower IGD in more number of runs compared to other algorithms, then followed by NSGA II, MOGA, MOACO and MOPSO algorithms. The below [Fig. 2] shows the result of IGD metric performance of the five algorithms. From the [Fig. 2] it is observed that when the number of cities is increasing, the IGD of NSGA, MOGA, MOACO and MOPAS algorithms are also increasing and the starting value of IGD is also very high. But in MOEA/D the IGD value is zero and when the number of cites is increasing value of IGD is going in a constant way. So from the definition of IGD, it is known that MOEA/D is having good diversity and convergence. All the other four algorithms are lacking in their performance with respect to IGD. So with respect to IGD, MOEA/D is giving best performance.

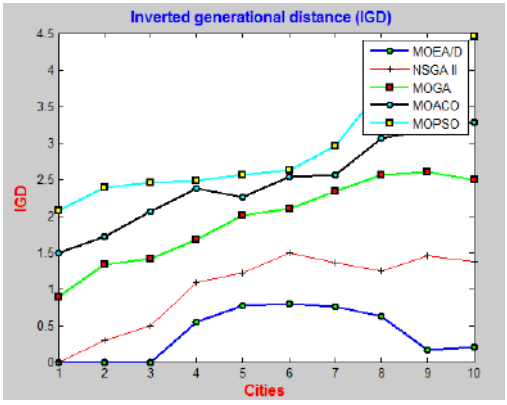


Fig. 2: IGD of MOEA algorithms.

The Average Computational Time (ACT) is also calculated to find the best algorithm among the five to solve MOTSP with respect to time complexity. The ACT of the MOEA are listed in the below [Table 2]. ACT is measure in seconds.

IIOAB JOURNAL
ISSN: 0976-3104

Table 2: Average Computational time (ACT) *(Sec)* of MOEA Algorithms

RGC/Algorithm variants	MOEA/D	NSGA II	MOGA	MOACO	MOPSO
1	5.63	15.31	18.80	23.75	26.83
2	7.05	23.53	25.45	35.68	26.04
3	17.09	61.52	50.28	56.57	27.41
4	34.23	60.73	102.28	134.85	179.02
5	34.24	93.42	105.78	170.47	228.61
6	38.83	120.38	170.03	230.78	240.99
7	41.44	196.29	240.09	399.76	503.51
8	89.21	266.27	450.34	630.54	782.52
9	134.92	747.23	876.22	934.23	1027.34
10	484.71	958.91	1021.34	1194.25	1324.86

For the [Table 2], MOEA/D is having the minimum computational time. NSGA II algorithm took almost three times more ACT than MOEA/D. Other algorithms are giving even higher time complexity. The ACT result of the algorithms is shown in the [Fig. 3]. When the number of cities is increasing the ACT also increases. MOEA/D has the lower computational time compared to other algorithms. So with respect to ACT also, MOEA/D is the best algorithm to solve MOTSP.

From the result analysis, MOEA/D is the best algorithm to solve MOTSP in all aspects. NSGA II is also giving good performance only but it is good when the number of objectives considered is only two [11] [13]. MOGA is giving good performance for stable environment but not for dynamic environment and the local search ability is also not good [35] [13].

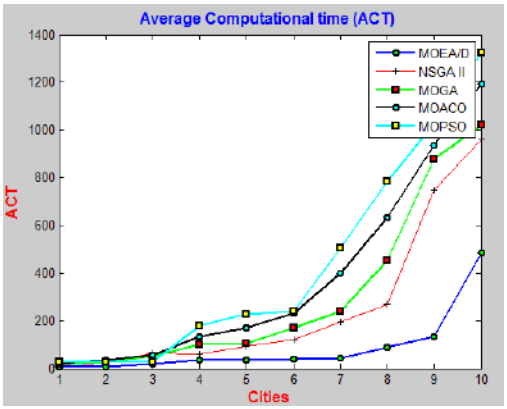


Fig. 3: ACT (sec) of MOEA Algorithms.

MOACO is performing well in dynamic environment but in ACO starting and destination nodes should be defined at earlier. MOPSO is good in global search but the convergence and local search are not at the expected level to solve MOTSP. MOEA/D is giving better performance in all aspect compared to other algorithms which is proved from the IGD and ACT analyses of the algorithms [18] [11] [17] [16].

CONCLUSION

In this paper the problem of TSP is explained in order to understand the concept and need of MOTSP which optimizes more objectives to find the best solutions. There are many methods to solve the MOTSP. Among those Multi-Objective Evolutionary Algorithms are the best methods to solve it efficiently. By evolving a population of solutions, multi-objective evolutionary algorithms (MOEAs) are able to approximate the Pareto optimal set in a single run. It results in nearest optimal solution within a reasonable time by optimizing many objectives simultaneously. The MOEA algorithms used to solve MOTSP are explained with its algorithm framework in section II. Each and every algorithm is best suited in some situation or environment to solve it. Among listed MOEA algorithms MOEA/D is giving the best solution for the MOTSP.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

This work is a part of the Research Projects sponsored by Visvesvaraya Ph.D. Scheme for Electronics & IT, Ministry of Electronics & Information Technology, India, and Reference No: PHD-MLA-4(44)/2015-16, dated August 2015. The authors would like to express their thanks for the financial supports offered by the Sponsored Agency.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Saraei M, Analouel R, Mansourl P. [2015] Solving of Travelling Salesman Problem using Firefly Algorithm with Greedy Approach, (June).

[2] Khanra A, Maiti MK, Maiti M. [2015] Profit maximization of TSP through a hybrid algorithm. Computers and Industrial Engineering, 88:229–236. https://doi.org/10.1016/j.cie.2015.06.018

[3] Fdhila R. [2014] Distributed MOPSO with dynamic Pareto Front driven population analysis for TSP problem, 294–299.

[4] Changdar C, Mahapatra GS, Kumar Pal. [2014] An efficient genetic algorithm for multi-objective solid travelling salesman problem under fuzziness. Swarm and Evolutionary Computation, 15:27–37. https://doi.org/10.1016/j.swevo.2013.11.001

[5] Lopes CR. [2015] Using Genetic Algorithms to minimize the distance and balance the routes for the multiple Traveling Salesman Problem, 3171–3178.

[6] Ruben Ivan Bolapos, Mauricio Granada Echeverry, John Willmer Escobar. [2015] A multiobjective non-dominated sorting genetic algorithm [NSGA-II] for the Multiple Traveling Salesman Problem, Decision Science Letters, 4: 559–568. https://doi.org/10.5267/j.dsl.2015.5.003

[7] Segura C, Coello CA, Miranda G, Leon C. [2017] Using multi-objective evolutionary algorithms for single-objective constrained and unconstrained optimization. Annals of Operations Research, 240(1): 217–250. https://doi.org/10.1007/s10479-015-2017-z

[8] Xu M, Li S, Guo J. [2017] Optimization of Multiple Traveling Salesman Problem Based on Simulated Annealing Genetic Algorithm, 25.

[9] Sidoti D, Avvari GV, Mishra M, Zhang L, Nadella BK, Peak JE, Pattipati KR. [2016] A Multiobjective Path-Planning Algorithm With Time Windows for Asset Routing in a Dynamic Weather-Impacted Environment. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 1–16. https://doi.org/10.1109/TSMC.2016.2573271

[10] Bazgan C, Gourvès L, Monnot J, Pascual F. [2013] Single approximation for the biobjective Max TSP. Theoretical Computer Science, 478, 41–50. https://doi.org/10.1016/j.tcs.2013.01.021

[11] Köksalan M, Tezcane Öztürk D. [2017] An evolutionary approach to generalized biobjective traveling salesperson problem. Computers and Operations Research, 79: 304–313. https://doi.org/10.1016/j.cor.2016.04.027

[12] Li M, Yang S, Liu X. [2015] Bi-goal evolution for many-objective optimization problems. Artificial Intelligence, 228: 45–65. https://doi.org/10.1016/j.artint.2015.06.007

[13] Rahimi M, Baboli A. [2014] A bi-objective inventory routing problem by considering customer satisfaction level in context of perishable product.

[14] Vachhani VL, Prajapati HB. [2015] Survey of Multi Objective Evolutionary Algorithms.

[15] Jiang J, Gee SB, Arokiasami, WA, Tan KC. [2014] Solving Vehicle Routing Problem with Stochastic Demand Using Multi-objective Evolutionary Algorithm, 1. https://doi.org/10.1109/ISCMI.2014.18

[16] Zhou A, Gao F, Zhang G. [2013] A decomposition based estimation of distribution algorithm for multiobjective traveling salesman problems. Computers and Mathematics with Applications, 66(10):1857–1868. https://doi.org/10.1016/j.camwa.2013.05.031

[17] Ke L, Zhang Q, Battiti R. [2013] MOEA/D-ACO: A multobjective evolutionary algorithm using decomposition and AntColony. IEEE Transactions on Cybernetics, 43(6): 1845–1859. https://doi.org/10.1109/TSMCB.2012.2231860

[18] Ke L, Zhang Q, Battiti R. [2014] Hybridization of decomposition and local search for multiobjective optimization. IEEE Transactions on Cybernetics, 44(10): 1808–1820. https://doi.org/10.1109/TCYB.2013.2295886

[19] Souza MZ De, Trinidad A Pozo R. [2014] A GPU Implementation of MOEA / D-ACO for the Multiobjective Traveling Salesman Problem, 6–11. https://doi.org/10.1109/BRACIS.2014.65

[20] GAO F, Zhou A, Zhang G. [2012] An Estimation of Distribution Algorithm based on Decomposition for the Multiobjective TSP, (Icnc), 817–821.

[21] Shim VA. (2012). A Hybrid Estimation of Distribution Algorithm for Solving the Multi-objective Multiple Traveling Salesman Problem, 10–15.

[22] Li W. [2012] A Parallel Procedure for Dynamic Multi-objective TSP. https://doi.org/10.1109/ISPA.2012.10

[23] Xing H, Wang Z, Li T, Li H, Qu R. [2017] An improved MOEA/D algorithm for multi-objective multicast routing with network coding. Applied Soft Computing Journal, 59:88–103. https://doi.org/10.1016/j.asoc.2017.05.033

[24] Kuo RJ. [2017] A Fuzzy Multi-Objective Vehicle Routing Problem for Perishable Products Using Gradient Evolution Algorithm.

[25] Psychas ID, Delimpasi E, Marinakis Y. [2015] Hybrid evolutionary algorithms for the Multiobjective Traveling Salesman Problem. Expert Systems with Applications, 42(22):8956–8970. https://doi.org/10.1016/j.eswa.2015.07.051

[26] Psychas ID, Delimpasi E, Marinakis Y. [2015] Hybrid evolutionary algorithms for the Multiobjective Traveling Salesman Problem. Expert Systems with Applications, 42(22):8956–8970. https://doi.org/10.1016/j.eswa.2015.07.051

[27] Psychas ID, Marinaki M, Marinakis Y, Migdalas A. [2016] Non-dominated sorting differential evolution algorithm for the minimization of route based fuel consumption multiobjective vehicle routing problems. Energy Systems. https://doi.org/10.1007/s12667-016-0209-5

[28] Luo Y, Liu M, Hao Z, Liu D. [2014] An Improved NSGA-II Algorithm for Multi-objective Traveling Salesman Problem, 12(6):4413–4418. https://doi.org/10.11591/telkomnika.v12i6.5476

[29] GarciaNájera A, Bullinaria JA, Gutiérrez-Andrade MA. [2015] An evolutionary approach for multi-objective vehicle routing problems with backhauls. Computers & Industrial Engineering, 81, 90–108. https://doi.org/10.1016/j.cie.2014.12.029

[30] Bederina H. [n.d.] Evolutionary Multi-Objective Optimization Approach for the Vehicle Routing Problem with Uncertain Travel Time.

[31] Mathew N, Smith SL, Waslander SL. [2015] Planning Paths for Package Delivery in Heterogeneous Multirobot Teams. IEEE Transactions on Automation Science and Engineering, 12(4): 1298–1308. https://doi.org/10.1109/TASE.2015.2461213

[32] Maity S, Roy A, Maiti M. [2017] Journal of Industrial Information Integration An intelligent hybrid algorithm for 4- dimensional TSP. Journal of Industrial Information Integration, 5:39–50. https://doi.org/10.1016/j.jii.2017.02.001

[33] Pierre DM, Zakaria N. [2017] Stochastic partially optimized cyclic shift crossover for multi-objective genetic algorithms for the vehicle routing problem with time-windows. Applied Soft Computing Journal, 52:863–876. https://doi.org/10.1016/j.asoc.2016.09.039

[34] Riquelme N, Von LC. [2015] Performance metrics in multi-objective optimization, 1.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A STATE OF ART APPROACHES ON ENERGY EFFICIENT ROUTING PROTOCOLS IN MOBILE WIRELESS SENSOR NETWORKS

G. Kadiravan[1*], Pothula Sujatha[1], J. Amudhavel[2]

[1]Department of Computer Science, Pondicherry University, Puducherry, INDIA
[2]Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA

ABSTRACT

Background: *WSN is composed of large number of sensor nodes which sense the environmental parameter and forwards the data to sink node. The nodes near the sink in static WSN die earlier than other nodes (hot spot problem).* ***Methods:*** *Energy efficient routing protocols play a vital role in WSN. To eliminate hot spot problem, sink mobility is introduced. The mobile sink moves and receives data from sensor node using 1-hop communication and minimizes the energy dissipation of the sensor nodes and eliminates the hot spot problem.* ***Results:*** *In this paper, various type of routing techniques in mobile wireless sensor network is presented.* ***Conclusions:*** *A detail comparison of various protocols is also given at the end of the paper. Finally some future directions for energy efficient routing in WSN are also presented.*

KEY WORDS
Energy efficiency
Mobility
Routing protocol
Wireless Sensor Network

Received: 5 June 2017
Accepted: 25 July 2017
Published: 16 Sept 2017

***Corresponding Author**
Email:
kathirkadir@gmail.com
Tel.:+91 9500717145

INTRODUCTION

Recent advancement in the field of IT and IC technology leads to the deployment of cheap and small size sensor nodes. WSN is commonly used in various fields from military to civilian applications. Energy efficiency is a main problem in WSN [1]. Routing techniques plays a crucial role in reducing the energy consumption. Clustering is the most popular energy efficient technique where a set of sensor nodes are joined to form a cluster. From each cluster, a cluster head (CH) is selected and the rest of the nodes are defined as cluster members. In those situations, the CHs near the BS burdened with heavy load and exhaust their energy quicker than the nodes located far from the sink. This is termed as hot spot problem and various techniques are proposed to eliminate it [2]. An efficient technique to eliminate hot spot problem is the introduction of mobility in WSN. By introducing mobility in WSN [3], sensor node or sink node can move to any part of the network to transmit or receive data. So, the energy consumption in the network is distributed and connectivity is also attained. Mobility in WSN can be helpful to collect data from spare and disconnected WSN [4]. The usage of mobility in WSN reduces the transmission distance which results to lesser energy consumption. The main benefit of these techniques is flexibility and scalability [5].The architecture of mobile WSN with and without clustering is shown in [Fig. 1] and [Fig. 2].

A major drawback of mobility is the enormous amount of packet loss because of changes in network topology and increased delay [6]. Though mobility provides good performance, mobile sink and mobile nodes are expensive compared to static node and static sink. Using multiple mobile sinks, the performance is increased significantly compared to static WSN [7]. A mobile sink node is the effective solution when the cost of the network is not an important issue. The movement of mobile sink leads to topology charges and affects routing performance [8]. Various routing protocol are developed for mobile WSN with mobile sink. In this paper, a review of various routing protocols in mobile WSN is given.

ENERGY EFFICIENT ROUTING PROTOCOLS

Termite Hill

Termite hill is developed to distribute the load in WSN to eliminate hot spot problem [9]. It is based on the idea of using one mobile sink which capable of moving without any limits. The hot spot problem is avoided which is caused by the static nodes located need the sinks. Termite hill is a bio-inspired algorithm derived by the behavior of termites. The method is employed in both static and mobile WSN and implemented in WSN hardware. From the result, it produces high throughput, reduce energy consumption compared to AODV [10], in term of various speed. The network lifetime is improved than static sinks.

Mobicluster

Mobicluster is an effective clustering protocol for mobile sinks which moves in a predictable path [11]. It is used to cover isolated nodes which cannot move in the network. The CHs need to communicate with rendezvous nodes and takes turn in communicating data. There are 5 steps in mobicluster: clustering, rendezvous node selection, CH attachment to rendezvous node, data aggregation and forwarding to rendezvous nodes and communication between to rendezvous nodes and mobile sinks. An algorithm is also used to produce the cluster of various sizes. As a result, energy consumption in the network is well balanced. To select rendezvous nodes, a new algorithm is given which leads to reduce collision and increased throughput. For minimizing network lifetime, CHs can be rotated when their energy level in reduced.

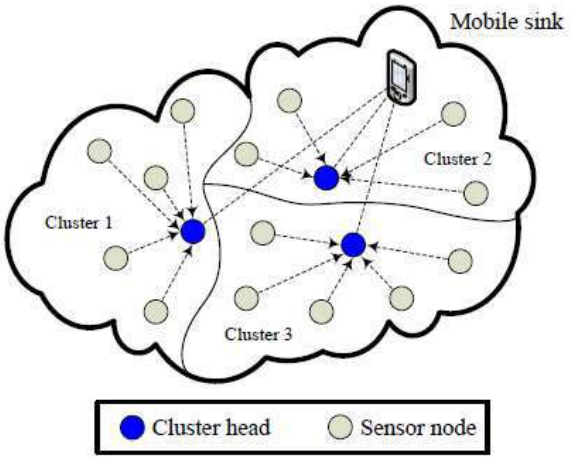


Fig. 1: Architecture of mobile WSN with clustering.

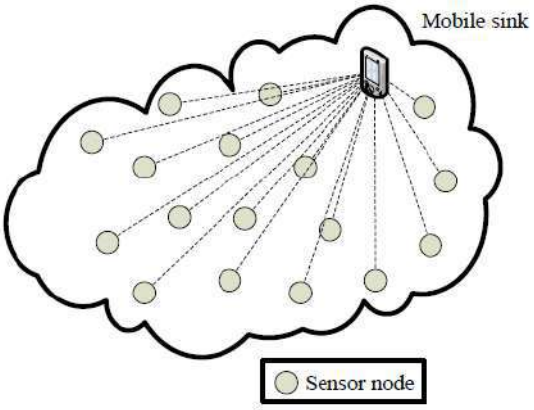


Fig. 2: Architecture of mobile WSN without cluster head.

Trace Announcing Routing Scheme (TARS)

Trace announcing routing scheme is developed to focus on various situations where the sink and targets need to be in mobility[12] .as both targets and sinks are mobile, a virtual grid based routing called TARS is developed. It is an improved version of target assisted routing scheme for WSN .it is based on the process capturing the mobile objects movement path by flooding and trace-announcing packet instead of path reconstruction. TARS maintain two tables: routing and tracking information. Additionally, a simple shortcutting method is also developed to reduce energy consumption.

W-L

W-L is an effective distance aware routing protocol with various mobile sinks [13]. To minimize the energy consumption, first order radio model is used to very transmission power with respect to the distance. The reduction in transmission energy reduces the interference. The energy dissipation at the transmitter (E_{TX}) and receiver (E_{RX}) with distance d for transmitting an l-bit data packet is computed in Eq. (1) and (2):

$$E_{TX}(l,d) = \begin{cases} l \times E_{elec} + l \times \varepsilon_{fs} \times d^2 if d \leq d_0 \\ l \times E_{elec} + l \times \varepsilon_{mp} \times d^4 if d > d_0 \end{cases} \quad (1)$$

$$E_{RX}(l) = l \times E_{elec} \quad (2)$$

where E_{elec} is the dissipated energy in transmitter or receiver and it is based on various factors like digital coding, modulation, filtering, and spreading of the signal. The distance threshold is defined as $d_0 = \sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$. Based on the transmission distance d, the free space (ε_{fs}) or multipath fading channel (ε_{mp}) is used for the transmitter amplifier. A relay node is chosen with higher energy and lesser distance to mobile sink. A parking position is used where the mobile sink can gather data. Here, the collection of data is not possible when the sink is in mobility .the performance is improved with multiple mobile sink .the drawback is the cost of mobile sink.

IIOAB JOURNAL
ISSN: 0976-3104

HARP (Hierarchical and Adaptive Reliable Routing Protocol)

HARP is a heterogeneous network based protocol which divides the nodes in two ways: regular nods and cluster nodes based on the residual energy [14]. CH section is done based on the remaining energy of the node. It constructs a hierarchical tree with two divisions: intra cluster and inter cluster. A mechanism to recover nodes and mobility management is introduced to reconstruct tree when link failure occurs .HARP is more energy efficient, reliable and scalable when compared to LEACH.

RAHMoN: (Routing Algorithm for Heterogeneous Mobile Network)

It splits the sensor nodes into static nodes and mobile proposed [15]. The energy of the static node is less and mobile node is of higher energy. It operates in 3 phases: network configuration, detection and selection of CH and data delivery to sink. It is assumed that every node can be selected as CH. the CH is selected based on residual energy, mobility level and distance to sink. It leads to effective routing in terms of less overhead and large number of data transmissions.

HSN (Heterogeneous Sensor Network)

HSN is clustering protocol with a mobile sink for heterogeneous WSN [16]. It partition the network in three ways based on the energy level. 1. H-nodes (higher energy level), 2.L-nodes (lower energy level), 3.sink (infinite energy level). H-nodes has longer data transmission range and high data rate compared to L-nodes, It produces better results than HARP and RAHMON, CH is stationary and 1-hop communication is provided. It uses PSO to adjust the movement of sink among CHs. It is useful for large scale WSN. It provides better results compared to static sink.

Clue Based Data Collection Routing Protocol

CBDCR [17] uses a mobile sink which moves in random paths instead of fixed path and it broadcasts it location information to a limited distance and does not broadcast to the entire network. The sensor nodes which receive the location information are known as watchers which can send or receive data and assume the hop(s) from the sensor node to the mobile sink. Then the watcher node saves the information as clue to the location of mobile sink for data transmission. When the mobile sink moves, the number of watchers is increasing and the data sensed by the nodes can be easily transmitted to the mobile sink based on these clues. Various simulations are done with mobile sinks in network to assess the performance of CBDCR and the results shows that CBDCR decreases the transmission of duplicate and balances the energy consumption of the network. The overview of CBDCR is shown in [Fig. 3].

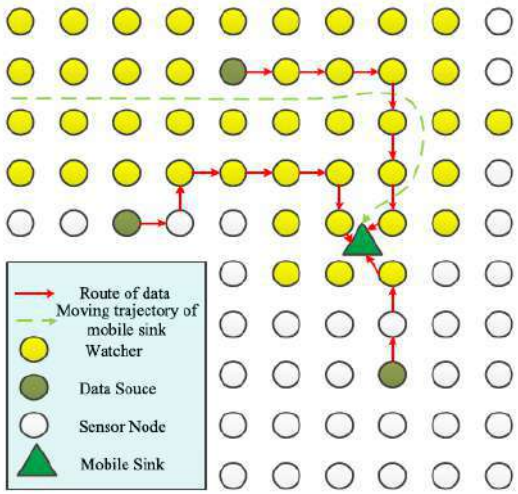


Fig. 3: Overview of clue based data collection routing protocol.

Zone based Energy Efficient routing Protocol

ZEEP is developed for both static and mobile nodes and needs no additional process for discovering path, maintaining routes or routing tables [18]. It employs the idea of dynamic forwarding and reduces the computation of the nodes. The simulations are done to verify that ZEEP achieves higher packet delivery ratio, reduced energy consumption by the network and results to maximum lifetime than well known protocol On Demand Distance Vector routing (AODV). ZEEP provides better performance than AODV protocol in terms of energy consumption and packet delivery ratio of the network. ZEEP, the customized version of ZBR also decreases the number of control packets created in the network compared to AODV and there is no need of path discovery or route maintenance. ZEEP is scalable as ZBR ZEEP can be easily extended to integrate optimizations which are default in AODV. A major advantage of ZEEP is simplicity.

Location aware sensor routing (LASeR)

Location aware sensor routing (LASeR) [19] protocol provides better solution to the issues of MWSNs. It concentrates on maximum reliability and minimum latency necessities of the emerging applications. It utilizes location information to retain a gradient field even in highly mobile environments, at the same time

as minimizing the routing overhead. This results to the usage of blind forwarding technique to circulate packets towards the sink. The protocol intrinsically uses multiple paths concurrently to generate route diversity and enhances its robustness. LASeR is proposed to employ in wide range of MWSN applications with independent land, sea or air vehicle. Logical expressions are derived and evaluated against the simulations. Extensive modeling and simulation of LASeR proves it is more flexible and robust. The results of LASeR is compared with advanced MWSN routing protocol includes the high performance mobility adaptive cross-layer routing protocol, as well as adhoc on-demand distance vector (AODV) and optimized link state routing. Protocols are analyzed based on the performance metrics such as packet delivery ratio, end-to-end delay, overhead, throughput and energy consumption. The performance of LASeR in several harsh environments shows that the proposed method is significantly better than the existing protocols. The packet structure of Laser is shown in [Fig. 4].

Field name	Node ID	Location	Data	Priority bit	Packet ID
Size (bits)	$\lceil \log_2 n \rceil$	$\left\lceil \log_2 \left(\frac{\sqrt{2} \cdot L}{Q_L} \right) \right\rceil$	L_{data}	1	$\lceil \log_2 n \rceil$
Total size (bits)	$L_p = 2 \cdot \lceil \log_2 n \rceil + \left\lceil \log_2 \left(\frac{\sqrt{2} \cdot L}{Q_L} \right) \right\rceil + L_{data} + 1$				

Fig. 4: Packet structure of LASeR

Table 1: Comparison of various energy efficient routing protocols

Protocol	Mobile element	Moving path	Node type	No. of sink	Application
Termite-hill	Sink	Random	Static node	Single	WSN with one mobile sink
TARS	Sink and targets	Random	Static node	Single	Location aware WSN
mobicluster	Sink	Fixed	Static node	Single	fixed paths of mobile sinks in WSN
W-L	Sink	Rectangle boundary	Static node	Single	Distance aware WSN
HARP	Sink	Random	CH and normal nodes	Single	Reliable WSN
RAHMON	Cluster based sink	Random	Mobile and static node	single	Hydropower plant
HSN	Sink	Random	l-node h-node sink	sink	Large scale WSN

COMPARISON AND DISCUSSION

In addition to eliminating hot spot problem, routing protocols are used in sparse and disconnected WSNs. The movement of sink or the sensor node reduces the distance of data transmission to preserve energy. This results to the uniform distribution of load among the nodes. The main benefit of these techniques is flexibility and scalability. Though these techniques have some advantages, severe drawbacks are also present. First, packet loss and latency is high due to topology changes in the network. Next, the mobile sinks are very expensive when compared to stationary nodes and the usage of mobile sink is doubtful. Using a number of mobile sinks in the network achieves greater performance than a single mobile sink, when the cost is not a consideration. When the cost is an important issue, selecting adequate number of mobile sink become an issue needs to be solved. The movement path of mobile nodes plays a major role on network topology and performance. MobiCluster and W-L follows a predefined path which is easier and adaptable to various situations. The nodes near the sink consume more energy than the nodes located away from the sink. This method does not equally distribute the load in the network. In Termite-hill and TARS, the sink changes the path based on the present network conditions without any limits. It is more convenient, but the implementation is very difficult and results to uncertainty. When the mobility speed on the nodes is low in large network, data latency will be increased significantly. Many researches need to done in the area of designing delay guaranteed routing protocols which is highly suitable for practical conditions. All the reviewed protocols used the same first order energy model, except W-L. As the routing protocols in homogeneous WSN eliminates hot spot problem, using mobile nodes in heterogeneous WSN also eliminates hot spot problem results to improved energy efficiency and uniform distribution of energy among nodes. [Table 1] compares the various reviewed protocols based on their characteristics.

IIOAB JOURNAL
ISSN: 0976-3104

CONCLUSION

Wireless sensor network (WSN) is composed of large number of sensor nodes which sense the environmental parameter and forwards the sensed data to sink node. The nodes near the sink in static WSN die earlier than other nodes (hot spot problem). Energy efficient routing protocols plays a vital role in WSN. To eliminate hot spot problem, sink mobility is introduced. The mobile sink moves and receives data from sensor node using 1-hop communication and minimizes the energy dissipation of the network and eliminates the hot spot problem. In this paper, various type of routing protocols in mobile WSN is presented. A detail comparison of various protocols is also given at the end of the paper. Finally some future directions for energy efficient routing in WSN are also presented. In future, different routing metrics like spatial reusability can be considered to improve the network throughput. Future work of the mobile routing protocols is network routing metrics: New routing metrics such as spatial reusability should be taken into consider in order improve the network throughput. Secure routing: Node and sink node act as perceived role and a router and makes it a vulnerable attack. Data rate consumes: To reduce high data rate in mobile sink and reduce energy consumption and transformation in multiple mobile sink.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Hiren Kumar, Deva Sarma, RajibMall, Avijit Kar. [2016] E2R2: Energy-Efficient and Reliable Routing for Mobile Wireless Sensor Networks. IEEE SYSTEMS JOURNAL, 10 (2).

[2] Jaichandran R, Irudhayara AA, Raja JE. [2010] Effective strategies and optimal solutions for hot spot problem in wireless sensor networks (WSN), Proc. IEEE 10th Int. Conf. Inf. Sci. Signal Process. Appl. (ISSPA). 389-392.

[3] Sara GS, Sridharan D. Routing in mobile wireless sensor network: A survey. Telecommun Syst. 57(1):51_79.

[4] Tunca C, Isik, Donmez MY, Ersoy C. [2014] Distributed mobile sink routing for wireless sensor networks: A survey. IEEE Commun. Surveys Tuts. 16(2):877-897.

[5] Yu S, Zhang B, Li C, Mouftah HT. [2014] Routing protocols for wireless sensor networks with mobile sinks: A survey. IEEE Commun. Mag. 52(7):150-157.

[6] Li, Zhang H, Hao B, Li J. [2011] A survey on routing protocols for large-scale wireless sensor networks. *Sensors*, 11(4):3498-3526.

[7] Basagni S, Carosi A, Melachrinoudis E, Petrioli C, Wang ZM. [2008] Controlled sink mobility for prolonging wireless sensor networks lifetime. *Wireless* Netw. 14(6):831-858.

[8] Tunca C, Isik S, Donmez MY, Ersoy C. [2014] Distributed mobile sink routing for wireless sensor networks: A survey. *IEEE Commun. Surveys Tut.* 16(2):877-897.

[9] Zungeru AM, Ang LM, Seng KP. [2012] Termite-hill: Routing towards a mobile sink for improving network lifetime in wireless sensor networks.' in Proc Int Conf Intell Syst Modelling Simulation, 622-627.

[10] Perkins CE, Royer EM. [1999] Ad-hoc on-demand distance vector routing. in *Proc. IEEE WMCSA*, 90-100.

[11] Konstantopoulos C, Pantziou G, Gavalas, Mpitziopoulos A, Mamalis B. [2012] A rendezvous-based approach enabling energy-efficient sensory data collection with mobile sinks.' *IEEE Trans. Parallel Distrib. Syst. 23(5):*809_817.

[12] Chi YP, Chang HP. [2012] TARS: An energy-efficient routing scheme for wireless sensor networks with mobile sinks and targets,' in Proc. IEEE Int. Conf. Adv. Inf. Netw. Appl.128-135.

[13] Wang J, Li B, Xia F, Kim CS, Kim JU. [2014] An energy efficient distance-aware routing algorithm with multiple mobile sinks for wireless sensor networks. *Sensors*, 14(8):15163-15181

[14] Atero FJ, Vinagre JJ, Ramiro J, Wilby M. [2011] A low energy and adaptive routing architecture for efficient field monitoring in heterogeneous wireless sensor networks. in Proc. IEEE Int. Conf. High Perform. Comput. Simulation. 449-455.

[15] Vilela MA, Araujo RB. [2012] RAHMoN: Routing algorithm for heterogeneous mobile networks. in *Proc. 2nd Brazilian Conf. Critical Embedded Syst. (CBSEC)*, 24-29.

[16] Sudarmani R, Kumar KRS [2013] Particle swarm optimization-based routing protocol for clustered heterogeneous sensor networks with mobile sink. Amer J Appl Sci 10(3): 259-269.

[17] Guisong Yang, Huifen Xu, Xingyu He, Liping Gao, Yishuang Geng, Zhunxue Wu. A Clue Based Data Collection Routing Protocol for Mobile Sensor Networks. in Digital Object Identifier 10.*1109/ACCESS.2016.2635697.*

[18] Juhi R Srivastava, Sudarshan TSP. [2013] ZEEP: Zone based Energy Efficient Routing Protocol for Mobile Sensor Networks. in 978-1-4673-6217-7/13/$31.00 ©2013 IEEE

[19] Tom Hayes, Falah H Ali. [2016] Location aware sensor routing protocol for mobile wireless sensor networks in IET Wirel. Sens. Syst., 1-9.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A STUDY ON IDENTIFICATION OF STATIC AND DYNAMIC PROTEIN COMPLEX AND FUNCTIONAL MODULE IN PPI NETWORK

Subham Datta[1*], Dinesh Karunanidy[1], J. Amudhavel[2], Thamizh Selvam Datchinamurthy[1], Subramanian Ramalingam[1]

[1]Department of Computer Science and Engineering, Pondicherry University, Puducherry, INDIA

[2]Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA

ABSTRACT

The availability of large-scale high-throughput experimental data has provided the opportunity of exploring biological networks to reveal complex structure and organization in the cell. Proteins are the main actors to perform cellular functions and majority of them do not carry out their functions in isolation but by interacting together in a complex manner. Protein complexes and functional modules are two important constructs formed by physical interactions among proteins. Identifying protein complexes and functional modules are crucial to understand the principles of cellular organization with important applications in disease diagnosis and therapy. However, experimental detection of protein complexes and functional modules is highly limited in the current state-of-the-art high-throughput experimental techniques. Thus computational approaches for detecting protein complexes and functional modules from protein-protein interaction data are valuable complements to the experimental techniques. From computational viewpoint, mining objects like protein complexes and functional modules from protein-protein interaction data (equivalently from protein-protein interaction networks) is computationally challenging task, since many problems related to determining structural properties of graphs are often NP-hard in nature. In this survey, we review the current state-of-the-art as computational techniques as well as recent emerging techniques for detecting protein complexes and functional modules, and discuss some promising research directions.

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

IIOAB JOURNAL
ISSN: 0976-3104

Retracted by authors on 22 July, 2019 due to plagiarism and inappropriate information and citations

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A NOVEL JAVA MACAQUE ALGORIHTM FOR TRAVELLING SALESMAN PROLEM

Dinesh Karunanidy[1]*, J. Amudhavel[2], Thamizh Selvam Datchinamurthy[1], Subramanian Ramalingam[1]

[1]*Department of Computer Science and Engineering, Pondicherry University, Puducherry, INDIA*

[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT

Abstract: *The natural evolution is originated from the principle of nature and it is completely based on the Charles Darwin theory "survival of the fittest". The naturally inspired algorithms are efficiently used for solving the real-world problems are motivating. This paper proposes the new java macaque algorithm based on the social behavior of the java macaque monkey. The behavior of the java macaque monkey is analyzed differently from the existing algorithm. Hence, the Java Macaque algorithm has the ability to solve the real-world optimization in an efficient way. In order to verify the efficiency of the proposed algorithm, the experimentation was performed with standard test- bed of Travelling Salesman Problem (TSP). Experimentation results clearly illustrate the consistency of the proposed algorithm over the existing algorithm like genetic algorithm (GA), particle swarm optimization (PSO) and ant colony optimization (ACO).*

KEY WORDS
Natural-inspired algorithm, JMA, java macaque, optimization algorithm, TSP, GA, PSO, ACO.

Received: 27 June 2017
Accepted: 28 July 2017
Published: 17 Sept 2017

***Corresponding Author**
Email:
dineshhumar@gmai.com
Tel.: +91 8056472815

INTRODUCTION

Natural-inspired Optimization Algorithms are become quite popular because of its simplicity and flexibility for solving large-scale optimization problem [1]. Most optimization problem remains unsolved or poorly solved, by the exact solving methods. In the last few decades, natural computing [2] has stood out as the essential technique for solving the complex real-world problems. Generally, the Natural-inspired algorithm [3,4] is the combination of the imitated process in nature and also inspired from nature. The Natural-inspired optimization algorithms are broadly divided into two categories such as Evolutionary Algorithms (EAs) and Swarm Intelligence (SI). The evolutionary algorithms are effective for solving the discrete optimization problem whereas the continuous problems are solved by swarm intelligence. But the wide optimization problems (M. Wagner, 2013) are intriguing because the main objective is to find the optimal arrangement, ordering or selection process. In particular, the famous evolutionary algorithms are the genetic algorithm, genetic programming, evolutionary strategies and evolutionary programming [5,6,7,8]. Similarly, the popular swarm intelligences are ant colony optimization [9], particle swarm optimization [10], clonal selection algorithm [11], cuckoo search [12], bat algorithm [13] etc.

RELATED WORK

In the literature, there are several algorithms for solving the combinatorial optimization problem. But the most dominant optimization algorithm is the genetic algorithm and particle swarm optimization. The Genetic algorithm is introduced by John Holland [5] which operates on the behavior of evolution process. The basic unit of the genetic algorithm is called as chromosomes or genotype of the genome that can be an optimal solution to the problem called phenotype. However, the primary contrast of the GA is crossover and mutation which play the vital role in exploitation and exploration of the search process. The PSO was introduced by Kennedy and Eberhart [14]. The basic principle of PSO is "collective intelligence" to find the candidate solution in the search space. It is also population (called swarm) based algorithm with the basic entity as particle [15,16]. Both the GA and PSO are widely used in the literature for solving the optimization problem [17,18, 19].

There are many natural inspired optimization algorithms [20] which are inspired by natural evolution and swarm optimization. Similarly, the social behavior of the monkey is also considered as the important natural inspired algorithm. Zhao and Tang [21] introduced the monkey algorithm (MA) based on the mountain-climbing process of monkeys. The main process of MA is climb-process, watch-jump process and somersault process. In the monkey algorithm, the random process is used for generating the initialization process which is followed by the climbing process used to search for the local optimal solution. Finally, the somersault process is employed for updating the current position. Zheng [22] has proposed the improved monkey algorithm with chaotic search method for initialization process and also introduced two parameters such as evolutionary speed factor and aggregation degree. The evolutionary speed factor is utilized to dynamically change the step length of each monkey in climb process whereas the aggregation degree dynamically modifies the eyesight and somersault distance of monkeys.

J.C. Bansal et al. [23] proposed Spider Monkey Algorithm (SMO) based on the promising behavior of monkey which holds well for the local search space problem. Sandeep et al suggested the fitness based position update for the Spider Monkey Algorithm that enriches the convergence rate. Even though, the algorithm shows its significance in local search process but lacks in the global search process. Menga and Pana [24] have proposed the Monkey King Evolution (MKE) algorithm based on the action of monkey king in the Chinese mythological novel. MKE algorithm is based on the transformation of monkey king into the

small monkey to solve the tough problem. The solution of the small monkey can be given as feedback solution to the monkey king. The basic ideology for exploitation is achieved by small monkey whereas the remaining monkeys in the population are used for exploration. Hence, this motivates our research to develop the natural- inspired algorithm which adopts well for both local and global search process.

SOCIAL BEHAVIOUR OF JAVA MACAQUE MONKEY

In this section, section 2.1 explains the Social behavior of Java Macaque Monkey and section 2.2 explains about our Proposed Java Macaque Monkey Algorithm.

Java Macaque Monkey is an important primitive which lives in the social structure. The male and female java macaque are shown in [Fig.1] and [Fig. 2]. and it has 95% gene similarity when compared with human being [25]. The important characteristics of the human such as joy, fear, violent, loyal and dedication are often seen in java macaque are shown in [Fig.3, 4, 5,6 and 7].

Fig. 1: Male with moustache.

Fig. 2: Female with beard.

It follows the social hierarchy structure among the java macaque individuals. The high-ranking java macaque dominates the low order JM individuals for accessing the food and resources. Usually, the java macaque lives in a group of 20-60 individuals with the average lifespan of 40 years [26]. There are many groups that found in the same environment but each group tries to dominate over their region of living.

Fig. 3: Joy.

Fig. 4: Fear.

Fig. 5: Violent.

Fig. 6: Loyal. **Fig. 7:** Dedication to offspring.

Fig. 8: Group behavior of java macaque monkey.

[Fig. 8] clearly demonstrate the group behavior of java macaque monkeys and each group has their social hierarchy among the individuals [26]. The dominant individual of the group is called as “Alpha Male” which also means as the leader of the group. Each group is majorly subdivided into two division such as dominant and non-dominant. Dominants are the higher rank individuals in the population which lives at the top of the social hierarchy while the Non-dominant are at the bottom of the social hierarchy. It is noteworthy that the advantages of the dominant are more access to resources, high power and social protection when compared with another non-dominant individual. So, the non-dominant has only less access to food and other resources. But apart from the dominant and non-dominant, in general, the male java macaque dominant over the female individuals. The learning method of java macaque is depending

on the environments circumstances. [Fig.9, 10] shows the social learning of individuals such as playing, Fighting and learning from the elders.

Fig. 9: Playing.

Fig. 10: Fighting.

The next important stages of JM culture are the reproduction. The reproduction is always depending on the social hierarchy of the individuals (i.e.) the male or female JM selects their partner based on the social rank [27]. Especially, during the reproduction time, the male java macaque sends the secret signal towards the female JM by making special noise and raising the eyebrow as presented in [Fig.11].

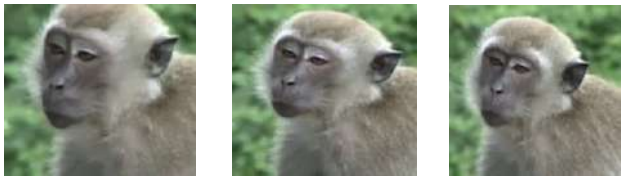


Fig. 11: Raising the eye brow.

In every group, the number of female individuals is higher in comparison with male (http://animaldiversity.org/accounts/Macaca_fascicularis/). So, each male has the relationship with the maximum of 3 female java macaque in the group. Then the new born java macaque has the social status depends on the parent's social ranks. In detail, the child born for the dominant rank individuals in the group has given importance when compared with individuals in the group [28]. In the same manner, the child of low-rank individuals is given less importance and consider as low-ranked individuals in the group. Generally, the children learn from the dominant female of the group. Thus, the newborn java macaque is protected by the other individuals from the group (i.e.) mother or other JM of the same social order as displayed in [Fig.12].

Fig. 12: Protection of new born java macaque monkey.

Engelhardt and Pfeifer [27] the major challenge for the alpha male is to protect their group and also find the suitable environment for finding the food and shelter. It also has to concentrate the behavior of the other individuals within the group and solve the dispute among the individuals as shown in [Fig.13]. The alpha male has more right to food resource of the group. The newborn male and female java macaque reach the sexual maturity in 4 and 3 years respectively. The male JM which reaches the sexual maturity has forced to leave the group and it is called as "stray male".

Fig. 13: Alpha Male solve the dispute among individuals.

The stray male has to search for another group for their protection. In order to join a new group, stray have only two ways (i.e.) either defeat the alpha male of the group or sexually attract the female JM of the group. If the stray male defeats the alpha male of the group then it became the new alpha male of the group and defeated alpha male has to leave the group and become stray male. But in most cases, the stray male attracts the female JM and that female JM convince other group members to add the stray male into the group. The fittest male java macaque has the right to become the alpha male of the group. The important factors which influence the dominance hierarchy of the java monkeys are age, size and Fighting

techniques. Hence, the lower ranking java macaque can become the higher ranking alpha male by improving the size and Fighting skills. The learning or cultural behavior is one of the special characters of the java macaque among other non-human primates. There are different types of learning pattern are followed among the java monkeys. Mostly the learning process is based on the circumstances of the environment and also the behavior of the other monkeys within the group. Especially, the grooming process is maintained among the female within the group as exposed in [Fig.14]. The lower ranking female monkeys are used to groom the higher ranking female java macaque in order to increase the access to food and protection.

The communication and perception of the java macaque genus are the facial expressions, vocalization, and body gesture. On the other hand, the chemical (olfactory) and physical signal are the main modes of communication among the java macaque monkeys.

Fig. 14: Grooming.

PROPOSED JAVA MACAQUE ALGORITHM

The Java Macaque Algorithm (JMA) was inspired by the social behavior of java monkeys which falls under the categories of Natural-inspired Algorithm. The intelligent and social behavior of the java macaque monkeys with multi-group population motivates the authors to develop new Java Macaque Algorithm. Hence, the Fig.15 clearly demonstrates the life-cycle of java macaque monkey suits well for solving the large-scale optimization and engineering problem. The proposed JMA is one among the naturally inspired algorithm with the feature like adaptive and self-organization. The java monkey exhibits adaptive social behavior which often responds according to the environmental changes and also has the dynamic process which maintains the global order emerging from the local interaction among the java monkeys in the group.

Java Macaque Algorithm (JMA)

Step 1: [Begin] Initialize n number of groups and individuals.
Step 2: [Evaluation] Evaluate the fitness value (i.e.) social rank of every individual in each group and fittest is “Alpha Male”.
Step 3: [Reproduction] Generate the new individuals for each group
  a. [Selection] select the two individuals from each group based on their social rank or fitness value.
  b. [Reproduction] create the new individual using the reproduction process (CROSSOVER).
Step 4: [Evaluation of Children] Evaluate the fitness value of newly generated individuals.
  a. [Stray Male] If the fitness value of the individuals is > 3/4 of Alpha male then become Stray male.
  b. [Female] Remaining individuals in children are female which stays in the same group.
Step 5: [Male Replacement] Stray Male find the suitable group using fitness value.
  a. If stray male fitness value > alpha male then become alpha male.
  b. Else if stray male fitness value > fitness value of the higher-ranking male then become higher ranking individual.
  c. Else if stray male is removed.
Step 6: [Learning] Improves the behavior of the individual from the surrounding and other individuals in the group.
Step 7: [Termination] Maintain the group size with new population and Repeat the above of the certain number of generation.

Initialization

Initialization is the important process of the java macaque algorithm which generate the ‘n’ number of groups and ‘m’ number individuals or macaque per group. The initialization process begins with the minimum number of individuals per group (i.e.) group size at initialization process is 20 individuals. But the maximum number of individuals for each group is 60.

Evaluation of fitness value

The fitness value of the individuals is calculated for all the individuals. The individuals with best fitness value for each group is called as “Alpha male”. The alpha male is the leader of the group which dominated

IIOAB JOURNAL
ISSN: 0976-3104

other individuals. The average fitness value of each group is utilized for dividing the group into two major subdivision such as “Dominant” and “Non-dominant”. The individuals which have fitness value above then the average fitness value is called as “Dominant”. The dominants are high fitness individuals or high social rank individuals of the group. On the other hand, the lower ranking individuals of the group are Non-dominant. Further, the dominant and non-dominant are classified into male and female based on the region fitness value using hashing method. From this region, the individual with maximum fitness is considered as male and remaining individuals are female. After this, the Average fitness value of dominant and non-dominant are calculated.

Reproduction

The new individuals are created by selecting the parent of same social rank. The parent selection procedure is based on the social hierarchy of the java macaque. Then the new individuals are generated by the combination of the both parent behavior. The probability of male to participate in reproduction is only 3% in one generation whereas the female has the probability of 1% respectively. The survival of the new offspring is based on the social ranking of the parent.

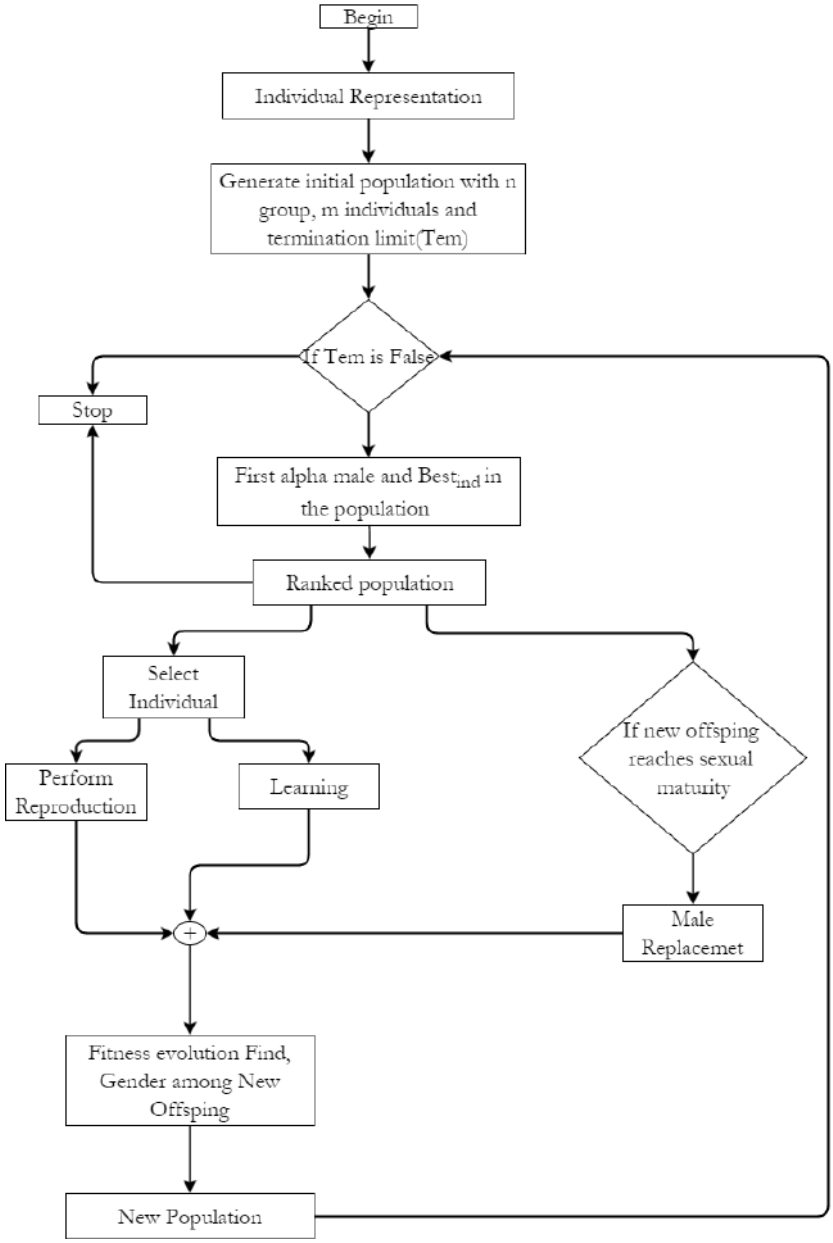


Fig. 15: Flowchart of Java Macaque Algorithm.

Fitness evaluation of new individuals

The newly generated individuals are evaluated using the fitness function. The fitness value of the male in the new offspring if greater than the 3/4 of Alpha male, then it became stray male. The female individuals

IIOAB JOURNAL
ISSN: 0976-3104

of the group are protected by the alpha male and other dominant individuals of the group but the stray males are sexually matured male which has been forced to leave the group.

Male replacement

The stray male has to find the appropriate group for leaving by using its fitness value. If the fitness value of the stray male is higher than the fitness of alpha male of any group, then it became the alpha male of the group Then if the fitness value matches the average fitness value of dominant of any group it becomes the dominant individual of the group. On the other hand, the stray male has the chance to become the non-dominant individual of the group if the fitness value is higher than the average fitness value of the non-dominant. In some cases, if the stray does not suit for any group has been eliminated from the population.

Learning

Learning is considered as the important stages of the newly generated individuals which learn from the surrounding and environment. The learning procedure of the newly generated individuals is mainly depending upon the behavior of the dominated individuals of the same group. The grooming is the special kind of learning activity which takes place between higher and lower ranking females.

Termination

The population of each group is maintained by eliminating the male and female of the group with the new population and the above process is repeated until the generation limit is reached. The initial population of the JMA is generated using the random population initialization technique. But the multi-group based population plays an important role in solving the travelling salesman problem efficiently. Further, the group is subdivided into male and female.

The individuals in the population are ranking according to the fitness value in the population. Then the social ranking based selection process is used for selecting parents for reproduction. Meanwhile, the new offspring which reaches the sexual maturity called 'stray male' are forced to leave the nodal group. Then the stray male has to find the suitable group using male replacement process. Finally, all the individuals in the group have undergone the learning process from the surrounding and environment.

RESULTS

The significance of the proposed Java Macaque algorithm is demonstrated using the travelling salesman problem (TSP). The Travelling salesman problem is one of the most important optimization problems for the researchers has been chosen as the test bed. Naturally, the TSP belong the class of NP-hard problem and act as the testbed for the new optimization algorithm. Hence, the instances eil51, st70, eil101 and ch130 are chosen as the test dataset which is obtained from the standard library of TSPLIB [24].

Therefore, the fitness function (Fit) for the proposed algorithm has been calculated using:

$$Fit = \min\left\{\left(\sum_{i=1}^{P} dist(C_i, C_{i+1})\right) + dist(C_p, C_1)\right\}$$

Whereas,

- P refers to the number of cities in the individual,
- $dist(C_i, C_{i+1})$ Refers the distance between two cities C_i and C_{i+1},
- $dist(C_p, C_1)$ Refers to the distance between last city and first city during return after the tour.

The standard experimental setup for the proposed JMA: the number of groups (n=5) with the population size of (m=60) and executed up to 100 generations and the initial population of each group were generated randomly. The performance of the proposed algorithm is compared with the Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). Each algorithm was run on each instance 25 times and hence the best among the 25 runs are considered for analysis and validation purposes.

Parameter for performance assessment

As stated earlier, the efficiency of the proposed self-organized GA is demonstrated against classical GA techniques with the help of a set of standard assessment factors such as convergence rate, error rate, average convergence rate and convergence diversity, which is briefed as follows:

Convergence rate: Convergence rate of an individual in the population is defined as the percentage of fitness value obtained by the individual in accordance with the optimal fitness value. It can be formulated as follows:

$$convergence\,rate\,(\%) = 1 - \frac{Fitness - optimal fitness}{optimal fitness} \times 100$$

IIOAB JOURNAL
ISSN: 0976-3104

Average convergence: Average convergence rate can be defined as the average percentage of fitness value of the individual w.r.t to the optimal fitness value. It can be expressed as follows:

$$Avg.conv.(\%) = 1 - \frac{averagefitness - optimalfitness}{optimalfitness} \times 100$$

Error rate: Error rate of an individual is defined as the percentage of the difference between fitness obtained by the individual and optimal fitness value. It can be given as:

$$Errorrate(\%) = \frac{Fitness - optimalfitness}{optimalfitness} \times 100$$

These above assessment criteria form a strong base for demonstrating the performance of the proposed JMA in the related domain and hence adopted by most of the researchers in the related domain. By following the same, this research also uses the same for validating the performance of the intended research.

DISCUSSION

The computational result of the experimentation is illustrated in [Table 1]. Thus, the experimentation is directed to evaluate the performance of the proposed JMA with another existing GA, PSO and ACO. For instance, the sample instance eil51 may be taken for discussion. From the [Table 1] the best convergence rate for GA is 81.54%, PSO is 77.58% and ACO 90.82% but the convergence rate of JMA is 100% of the

Table 1: Computation result for GA, PSO, ACO and JMA

S.No	TSP Instance	Technique	Optimum value	Fitness		Convergence rate (%)	Error rate (%)	Average Convergence (%)
				Best	Average			
1	eil51	GA	426	522.44	790.66	81.54	18.46	14.40
		PSO		549.09	686.44	77.58	22.42	38.86
		ACO		469.09	616.44	90.82	9.18	55.30
		JVM		426.00	485.84	100.00	0.00	85.95
2	pr76	GA	108159	131092.80	200955.99	82.51	17.49	14.20
		PSO		136082.80	175945.99	79.48	20.52	37.33
		ACO		121425.45	156599.77	89.07	10.93	55.21
		JVM		108159.00	127534.83	100.00	0.00	82.09
3	pr144	GA	58537	71188.74	111140.30	82.23	17.77	10.14
		PSO		75189.00	95839.30	77.85	22.15	36.28
		ACO		65421.63	85103.84	89.48	10.52	54.62
		JVM		59083.15	70469.88	99.08	0.92	79.61
4	tsp225	GA	3919	4944.16	7533.00	79.27	20.73	7.78
		PSO		5255.16	6494.82	74.57	25.43	34.27
		ACO		4410.68	5905.76	88.85	11.15	49.30
		JVM		3961.27	4867.03	98.93	1.07	75.81
5	fl417	GA	11861	15945.57	23287.99	74.38	25.62	3.66
		PSO		16857.57	19975.99	70.36	29.64	31.58
		ACO		13564.25	18228.38	87.44	12.56	46.32
		JVM		12064.19	14865.06	98.32	1.68	74.67
6	u724	GA	41910	58608.09	83013.84	71.51	28.49	1.92
		PSO		60518.09	72853.87	69.25	30.75	26.17
		ACO		48155.24	69071.94	87.03	12.97	35.19
		JVM		43423.92	56971.93	96.51	3.49	64.06

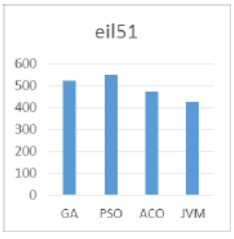


Fig. 16(a): eil51

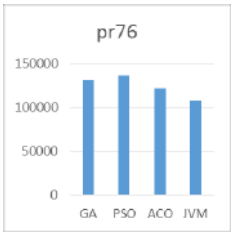


Fig. 16(b): pr76

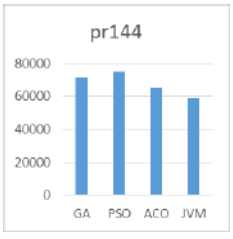


Fig. 16(c): pr144

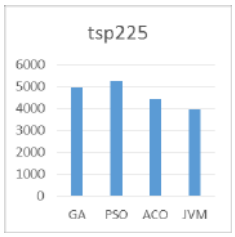


Fig. 16(d): tsp225

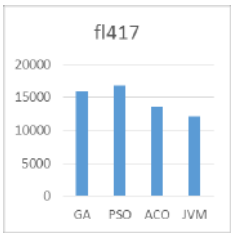


Fig. 16(e): fl41

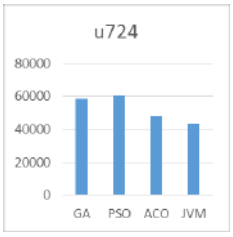
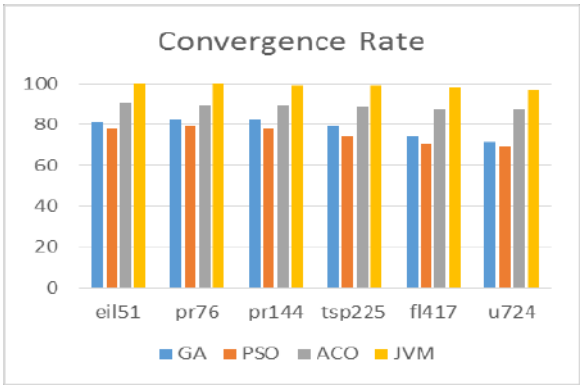


Fig. 16(f): u724

Fig. 16: Analyses the performance based on fitness value.



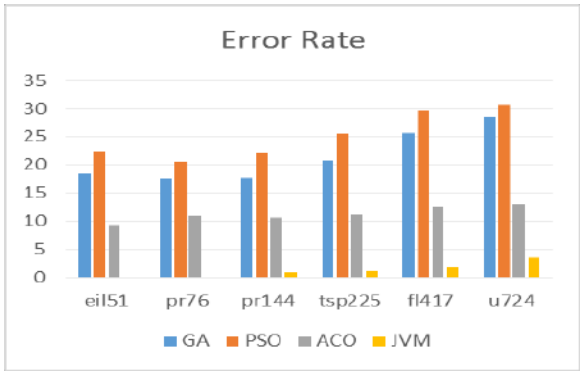


Fig. 17: Analyses the performance based convergence rate. **Fig. 18**: Analyses the performance based error rate.

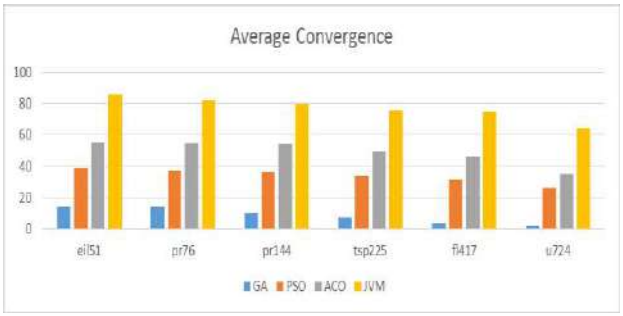


Fig. 19: Analyses the performance based average convergence rate.

same. Then the average convergence rate for JMA, GA, PSO and ACO are 85.95%, 14.40%, 38.86% and 55.30% respectively. Further, in terms of error rate also, this form of supremacy is continued for all instances. On the other hand, the convergence rate of the proposed algorithm is more than 96.51% for large instance u724, whereas the existing GA, PSO and ACO are achieved the rate of 71.51%, 69.25% and 87.03% for the same. It is exposed from the [Table 1] that the average convergence rate existing algorithm is lower than prosed JMA in all the instances. The examination of the proposed algorithm dominated the existing ACO, PSO and GA in all the performance assessment parameters.

Analyses based on fitness value

The fitness value is one of the important assessment criteria which give the tangible result of the optimal solution. From [Fig. 16], it is understood that the performance of java macaque algorithm outperforms the other existing algorithm in terms of fitness value. From the existing algorithm, the ACO suite well for solving the large-scale instance but the proposed JMA out-performed ACO in all the instances. Let us consider the sample instance name u724 in which the ACO has the best fitness value of 48155.24 whereas the JMA has the fitness value for the instance as 43423.92.

Analyses based on convergence rate

The convergence rate indicates the quality of the optimal solution generated from the population. [Fig. 17] illustrates the assessment of proposed algorithm against the existing algorithm w.r.t to the convergence rate. The convergence rate of the JMA is above 95% for all the instances. Hence the instance fl417 has

the lowest value of 70.36% convergence rate in the PSO and the highest of 98.32% for the JMA, while the ACO stand next to the proposed JMA with the value of 87.44%. The proposed java macaque algorithm dominates the existing algorithm in terms of convergence rate.

Analyses based on error rate

The performance of the algorithm w.r.t error rate is presented in [Fig. 18]. The evaluation of the proposed algorithm in terms of error rate is important for the analysis. The best error rate indicates how far the best individual convergence rate deviates from the optimal fitness value while the worst error rate demonstrates the difference between the convergence rate of the worst individual from the population and the optimal solution. Thus, the maximum value of error rate for the java macaque algorithm is 3.19%, Ant colony optimization is 12.97%, particle swarm optimization is 30.75% and the genetic algorithm is 28.49% correspondingly. Similarly, the minimum value of error rate of JMA is 0%, for instance eil51 and pr76 whereas the genetic algorithm has 18.46%, the particle swarm optimization has 22.42% and ACO has 9.18% for instance eil51 respectively. JMA obtained the better performance compared with existing algorithm.

Analyses based on average convergence rate

[Fig. 19] depicts the result analysis of average convergence rate in comparison of the proposed algorithm with existing algorithms. Thus, the investigation in terms of average convergence for JMA has shown the dominance over the other algorithm. On average, the average convergence rate of JMA has attained value above 60% but the GA, PSO and ACO, obtained value above 1%, 26% and 35%. The JMA achieves the maximum value of 85.95% while the ACO 55.30% respectively. By contrasting the performance of JMA and ACO for the average convergence rate is quite better for both algorithms.

CONCLUSION

In this paper, we proposed the new java macaque algorithm based on the social behavior of java macaque monkey. The features of the proposed JMA have been explained in this paper which suits well for solving the optimization problems. Hence, the authors have analyzed the performance of JMA on the NP-hard TSP problem and compared its results with the existing algorithm like GA, PSO and ACO. The robustness of the JMA has been clearly illustrated with respect to fitness value, convergence rate, error rate and average convergence rate. Therefore, it may be concluded that JMA is one of the efficient algorithms in the field natural-inspired optimization algorithms and achieves better result for the optimization problem.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
This work is a part of the Research Projects sponsored by Visvesvaraya Ph.D. Scheme for Electronics & IT, Ministry of Electronics & Information Technology, India, and Reference No: PHD-MLA-4(44)/2015-16, dated August 2015. The authors would like to express their thanks for the financial supports offered by the Sponsored Agency.

FINANCIAL DISCLOSURE
None

REFERENCES

[1] De Castro LN. [2006] Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications, Chapman & Hall/CRC. 267-323.
[2] De Castro LN. [2007] Fundamentals of natural computing: An overview, Physics of Life Reviews. 4:1-36.
[3] Chazelle B. [2009] Natural algorithms, ACM/SIAM Symp. Discret. Algorithms. 1:422–431.
[4] Ran Cheng P. [2016] Nature Inspired Optimization of Large.
[5] Holland J. [1975] Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Anbor.
[6] Goldberg DE. [1989] Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA.
[7] Man KF, Tang KS, Kwong S. [1996] Genetic algorithms: Concepts and applications, IEEE Trans. Ind. Electron. 43 (5):519–534.
[8] Flake GW. [1998] The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation, MIT Press, Cambridge, MA.
[9] Caro GDi. [2004] Ant Colony Optimization and its Application to Adaptive Routing in Telecommunication Networks', Intelligence. 41:374.
[10] Devarakonda S. [2012] Particle Swarm Optimization.
[11] Campus K, Engineering E, Campus K. [2012] Comparison Study For Clonal S Election. 4(4):107–118.
[12] Santos CAG, Freire PKMM, Mishra SK. [2012] Cuckoo search via levy flights for optimization of a physically-based runoff-erosion model, J Urban Environ. Eng. 6(2):123–131.
[13] Yang XS, Hossein Gandomi A. [2012] Bat algorithm: a novel approach for global engineering optimization. Engineering Computations, 29(5), pp.464-483.
[14] Kennedy J, Eberhart R. [1995] Particle swarm optimization, Neural Networks, Proceedings, IEEE Int. Conf. 4:1942–1948.
[15] Ben Guedria N. [2016] Improved accelerated PSO algorithm for mechanical engineering optimization problems, Appl. Soft Comput. J. 40:455–467.
[16] Ruan ZH, Yuan Y, Chen QX, Zhang CX, Shuai Y, Tan HP. [2016] A new multi-function global particle swarm optimization, Appl. Soft Comput. 49:279–291.
[17] Moon C, Kim J, Choi G, Seo Y. [2002] An efficient genetic algorithm for the traveling salesman problem with precedence constraints, Eur. J. Oper. Res. 140(3):606–617.

[18] Jayanthi M, Dinesh K, Sujatha P, Vengattaraman T, Dhavachelvan P, Subramanian R. [2017] GPS: A Constrain Based Gene Position Procurement in Chromosome for Solving Large Scale Multiobjective Multiple Knapsack Problems, Frontiers of Computer Science.
[19] Dinesh K, Amudhavel J, Rajakumar R, Dhavachelvan P, Subramanian R. [2017] A Novel Self-Organization Model for Improving the Performance of Permutation Coded Genetic Algorithm, Int. J. of Advanced Intelligence Paradigms.
[20] Chazelle B. [2009] Natural algorithms, ACM/SIAM Symp. Discret. Algorithms. 1:422–431.
[21] Zhao, R.Q. and Tang, W.S., [2008]. Monkey algorithm for global numerical optimization. Journal of Uncertain Systems, 2(3):165-176.
[22] Peng Z, Yin H, Pan A, Zhao Y. [2016] Chaotic Monkey Algorithm Based Optimal Sensor Placement. 9(1):423–434.
[23] Bansal JC, Sharma H, Jadon SS, Clerc M. [2004] Spider Monkey Optimization algorithm for numerical optimization, Memetic Comput. 6(1):31–47.
[24] Meng Z, Pan JS. [2015] Monkey King Evolution: A new memetic evolutionary algorithm and its application in vehicle fuel consumption optimization, Knowledge-Based Syst. 97:144–157.
[25] Van den Bercken JHL, Cools AR. [1980] Information-statistical analysis of factors determining ongoing behaviour and social interaction in Java monkeys (Macaca fascicularis), Anim. Behav., 28(1): 189–200.
[26] Veenema HC, Spruijt BM, Gispen WH, Van Hooff JARAM. [1997] Aging, dominance history, and social behaviour in Java-monkeys (Macaca fascicularis), Neurobiol. Aging, 18(5): 509–515.
[27] Engelhardt A, Pfeifer JB, Heistermann M, Niemitz C, van Hooff Ja Ra M, Hodges JK. [2004] Assessment of female reproductive status by male longtailed macaques, Macaca fascicularis, under natural conditions, Anim. Behav. 67(5):915–924.
[28] Dasser V. [1988] A social concept in Java monkeys, Anim. Behav. 36(1):225–230.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

REVENANT OF THE ECOSYSTEM: AN ENVIRONMENTAL BASED GREEN COMPUTING MODELS FOR VEHICULAR ROUTING PROBLEMS USING GENETIC ALGORITHM OPTIMIZATION APPROACH

M. Shanmugam[1*], J. Amudhavel[2]

[1]*Department of CSE, VFSTRU, Guntur, AP, INDIA*

[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT

Background: *Transportation plays a crucial role in our day to day life, without transportation facility it's really impossible to lead modern day routine life. The scope of transportation, supply chain management, logistics management plays a vital role within the delivery of products and services. The objective of this problem is to minimize the transportation cost and achieve efficient routing.* ***Methods:*** *We proposed an environmental oriented hybrid optimal routing algorithm for the road transportation system, we are looking this problem as a multi objective multi criteria because the goal is to minimize the distance and also the pollution.* ***Results:*** *The hybrid optimal based routing is achieved depending on the average convergence of distance and the pollution from the initial population.* ***Conclusions:*** *The experimental results evident that the new proposed technique performance well with the environmental factors. In addition to that, it is also outlined that further research work can be carried out to promote the proposed system with Vehicular Ad-hoc Network to provide betterment of Intelligent Transportation System.*

KEY WORDS
Genetic Algorithm, Routing, TSP, Transportation, *Optimization.*

Received: 23 June 2017
Accepted: 1 Aug 2017
Published: 17 Sept 2017

***Corresponding Author**
Email: shaninfo247@gmail.com
Tel.: +91-900-391-2326

INTRODUCTION

Vehicle Routing Problem is one of the most studied and well know problem. The complexity of this problem is providing efficient route for the fleet of vehicles. The formal factors affecting the routing strategy are time and distance bounds of the system. VRP is derived from the well know and stage Travelling Salesman Problem (TSP) [1]. There many optimization methods available to solve the NP hard problems, but some of the most complex asymmetric TSP [2] solved with the improved Genetic Algorithms (GA) optimization technique. GA employees three main process to formulate the optimal solution selection that is selection, cross over and mutation. Selection is one of the important modules and plays a vital role in optimal solution selection, in which the individuals are directly interconnected to their fitness value. If the fitness value is higher, then the chance of choosing the individual is higher. Next is cross over where the most typical solution is used. It takes place between the individual. The position of gene is chosen, where swapping operation may be carried out for possible set of solution. The point at which it is broken depends on the unusual selection of cross over point. This process represents the combinational operation of the individuals. The combination of both selection and cross over will generate the less quality solution. The last process of implementation is mutation. After the computation of selection and cross over, we may obtain a solution with same or different characteristics, done by the means of swapping operation (cross over) or by normally obtained one. [3] In this process, the possible setting of an individual has to be changed. Using this mutation alone, an unusual walk to search space has been generated. Generally mutation in GA will fine tune the optimal result. Our paper is organized as follows: Section 2 gives the fine-tuned literature survey and background information. In Section 3, we have discussed about the proposed system and algorithm, description for different scheme. In section 4 we will discuss about the experimentation methodologies and Section 5 describes about the experimental analysis of different scheme with different initialization techniques and their results. Finally Section 6 concludes the paper.

RELATED WORKS

Helsgaun, K et al [4] discussed Lin–Kernighan heuristic for symmetric TSP. [5,6] Introduced neighborhood based seeding technique to initialize the base population into GA for finding better route or individual. To improve the efficiency [7] introduced the new mutation operator to refine the TSP route. Heterogeneous Fleet Vehicle Routing Problem (HFVRP) is one of the common routing methods in VRP, which is based on different type of vehicle, capacities and the cost. The idea is to find the best route and to reduce the total distance. To solve this problem, [8] proposed a hybrid algorithm which consists of two methods, one is Local Search (LS) based heuristic and the second one is Set Partitioning (SP) formulation. While processing, the LS is used by the Mixer Integer Programming (MIP) solver to solve the SP model. This algorithm is only applicable for small scale, in that the efficiency is also less. By the influence of this problem same author proposed [9] Vehicle Routing Problems with homogeneous fleet, a parallel method that dynamically controls the dimension of the SP model. In [9] the LS based metaheuristic approach generates the routes with respect to a sequence of SP models with columns and analyzed that, this method is working properly for the large scale.

IIOAB JOURNAL
ISSN: 0976-3104

To solve the Multi-Objective Vehicle Routing Problems (MO-VRP) [10] proposed a fast approximation heuristics and the heuristic depends on the savings approach. The solutions are enhanced by the local search against the pare to-front in iterative process. Based on the savings heuristic the initial solutions are generated and the solution is approximated by pare to - front and then enhanced by the local search. This method has been tested on the beach mark it improves the initial approximation. [11] Proposed new microarray gene ordering strategy for solving combinatorial TSP optimization problem. [12,13,14] Introduced a hybrid particle swarm optimization algorithm for suggesting the best route for the fleet of vehicles, [15] discussed the adaptive neighborhood search method in VRP with time windows, [16] did the study with extended Variable Neighborhood Search problem (VNS) with Particle Swarm Optimization (PSO) and found that in context to time bound GA provide better results when compared to PSO.

The quality of the individual in the current generation is sent to the next generation, influenced by the Lamark's method. This method has been used to maintain the best solution throughout the process and many proposals are used to solve the application local search operator [17,18,19]. In [20] an enhanced the genetic operators (crossover and mutation) using the feasible solution and proposed an Improved Genetic Algorithm (IGA). To improve the effectiveness [7] [21], established three optimization strategies: immigration, local optimization and global optimization. Random population method is used, while generating the initial population, the chance of finding the optimal solution is very less and also the computation cost will be more.

In our survey we discussed, how the VRP is solved in using different methods, with different parameters and constrains like time windows constraints, multi-objective services, multi criteria vehicle routing [22,23,24]. We proposed a socially inspired transportation problem, which is based on the pollution in the path, we are routing the vehicle. The experiments are done using the slandered TSP bench mark instances [25] and then analyzed the performance with different Genetic Algorithm (GA) initialization techniques [26].

PROPOSED SYSTEM

In this work, the standard VRP is considered in a different perspective to propose a new environment concerned transportation problem in which the optimal path should be of least distance and also minimum air-pollution along the route. A pollution matrix of TSP, similar to distance matrix, is formulated to specify the pollution between each pair of cities. A pollution limit between the cities is the maximum allowed pollution value between any two cities in any feasible solution for the problem. During the formulation of solution, at each stage, inclusion of a new city is allowed only if the pollution value between the previous and the new city is less than that of maximum allowed pollution limit between the cities else, it would try to select the alternate city. The intelligent routing strategy for VRP in Hybrid Optimal routing can be represented as follows:

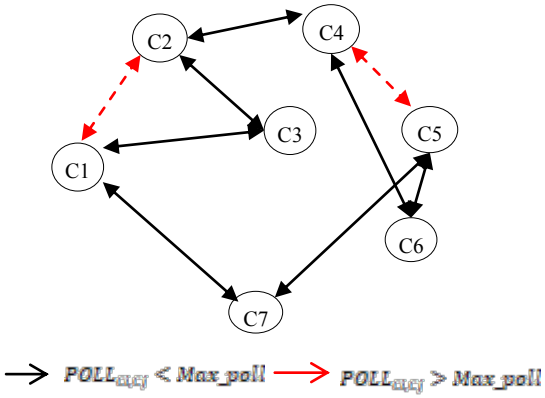


Fig.1: Sample intelligent routing strategy for VRP.

Let, the complete undirected graph $G = \{City_n, A\}$ and $DM(City_i, City_j)$ is the distance between the cities $City_i$ and $City_j$ such that $City_i \neq City_j$ and $DM(City_i, City_j) = DM(City_j, City_i)$.The pollution matrix for the TSP problem of size 'n'can be represented as $POLL(n \times n)$and $POLL(City_i, City_j)$ is the pollution between the $City_i$ and the $City_j$. In this proposal, the starting city remains same because the vehicle should start from a single predetermined source. The working principle of the proposed intelligent routing strategy with an example is illustrated below:

IIOAB JOURNAL
ISSN: 0976-3104

Algorithm

Hybrid Optimal based routing Algorithm (Pop, TP, Tc, G, n)

Step 1: Initialize $Gen \leftarrow 1, i \leftarrow 1, Size \leftarrow 1, TTC \leftarrow 0, TTP \leftarrow 0$

Step 2: Set optimal Distance, optimal Pollution and Maximum pollution $Opt\_Dist, Opt\_Poll \text{ and } Max\_Poll$

Step 3: Store the Population into a temporary variable, $CPop_{nxn} \leftarrow Pop_{nxn}$,

Step 4: Repeat through Step 13 **Until** $Gen \leq G$, Step 3 else go to Step 14

Step 5: Repeat through step 5.3 **Until** $l <= Popsize$, else goto Step 6

Step 5.1: $TCC \leftarrow (1 - (TC_l - Opt\_Dist)/Opt\_Dist) * 100$ //calculating the cost convergence of the individual

Step5.2: $TPC \leftarrow (1 - (TP_l - Opt\_Poll)/Opt\_Poll) * 100$//calculating the pollution convergence of the individual

Step5.3: $TradeOff\_Con_l \leftarrow TCC + TPC/2$// calculating the average of pollution and distance convergence the individual

Step 6: Select the best individual which is having maximum tradeoff Convergence and pass the best Individual to the next generation

Step 6.1: Repeat through Step 6.3 **Until** $i < ER$, else goto Step 7

Step 6.2: $position \leftarrow max(TradeOff\_Con)$// Position of the Individual with Maximum tradeoff convergence value will be acquired.

Step 6.3 $Pop_i \leftarrow CPop_{position}$// the individual in the position in temporary population is moved to the population

Step 7: Repeat through Step 8.6 **Until** $Size \leq PopSize$, else goto Step 9 $where\ ER < Size \leq PopSize$

Step 8: Choose the random parents Individuals, $P\_Indiv1 \text{ and } P\_Indiv2$

Step 8.1: Select the initial City $Init\_City, \text{ and } Length \leftarrow 1$, Size$\leftarrow$1

Step 8.2: $Indiv[Length] \leftarrow Init\_City$ // the first city in the parent individual is selected as initial city

Step 8.3: Repeat through Step 8.5 **Until** $Length \leq n$, else goto Step 5

$Cur\_City \leftarrow Indiv[Length]$ //the current city in the offspring

individual assigned as current city

Find the Position $Pos1 \text{ and } Pos2$ of the Current City in the Parent Individuals

$Pos1 \leftarrow find(P\_indiv1\,(Cur\_City)),\ Pos2 \leftarrow find(P\_indiv2\,(Cur\_City))$

In the [Fig. 1], the red dashed line shows the path between any cities, which current air-pollution is higher than the pollution limit of the problem and the black line shows the path with pollution within the limit. Assume that $C1$ is the starting city and the neighboring cities are organized in ascending order of their distance such that $(d(C1,C2) \leq d(C1,C3) \leq \cdots \leq d(C1,C7))$. The objective of the intelligent routing strategy for VRP in hybrid optimal routing is to choose the adjacent city devising lowest normalized value and the pollution between the cities are within the pollution limit has selected as next city. Starting from the city $C1$, the adjacent city is $C2$ with least normalized value but the pollution between the cities $C1$ and $C2$ exceeds the pollution limit. i.e. $POLL_{C1,C2} < Max\_poll$.

IF$Pos1 = 1,\ LLoc1 \leftarrow n$
Else IF$Pos1 = n,\ \ RLoc1 \leftarrow 1$
IF$Pos2 = 1$,$LLoc2 \leftarrow n$
Else IF$Pos2 = n$, $RLoc2 \leftarrow 1$
Evaluate the Tradeoff$Treadeoff_z$ from Previous City to Current City and Current City to Next City from the Parent Individuals using Normalization
$d_1 \leftarrow DM(P\_Indiv1(LLoc1), P\_Indiv1(Pos1))$
$d_2 \leftarrow DM(P\_Indiv1(Pos1), P\_Indiv1(RLoc1))$
$d_3 \leftarrow DM(P\_Indiv1(LLoc2), P\_Indiv1(Pos2))$
$d_4 \leftarrow DM(P\_Indiv1(Pos2), P\_Indiv1(RLoc2))$
$p_1 \leftarrow POLL(P\_Indiv1(Pos1-1), P\_Indiv1(Pos1))$
$p_2 \leftarrow POLL(P\_Indiv1(Pos1), P\_Indiv1(Pos1+1))$
$p_3 \leftarrow POLL(P\_Indiv2(Pos2-1), P\_Indiv2(Pos2))$
$p_4 \leftarrow POLL(P\_Indiv2(Pos2), P\_Indiv2(Pos2+1))$
$TC = \sum_{t=1}^{size(d)} {d_t}^2 , TP = \sum_{t=1}^{size(p)} {p_t}^2$ //sum of pollution and distance for the locations
$d'_t = \sum_{t=1}^{size(d)} d_t / TC$, $p'_t = \sum_{t=1}^{size(p)} p_t / TP$// normalizing the pollution and distance
$\forall [1, \leq z \leq Size(Loc)], \Omega_z \leftarrow TR * p'_z + (1-TR) * d'_z$

Step 8.4: Repeat through Step 8.6 Until$k < 4$, $else\ goto\ Step\ 9$ $where\ 0 < k \leq 4$

$Next\_City \leftarrow min(Tradeoff)$ //the location of the city with minimum cost will be acquired

Step 8.5: IF$Next_City \notin Indiv\ and\ POLL_{Cur_City, Next_City} < Max_Poll$, else goto Step 8.4

$Length \leftarrow Length + 1$,Update$Indiv[Length] \leftarrow next_City$

Step 8.6: $k \leftarrow k + 1$// increment the individual in the population

Step 9: Generate Random values $GeneC1, GeneC2$, $where\ 1 \leq GeneC1, GeneC2 \leq n$

$Swap\ \ Indiv(GeneC1) \leftarrow Indiv(GeneC2)$, $SwapIndiv(GeneC2) \leftarrow Indiv(GeneC1)$

Step 10:$Pop_{size} \leftarrow Indiv$, $Size \leftarrow Size + 1$ // move to the next individual

Step 11: Evaluate the cost of each Individual in the Population

$$\forall [1, \leq i \leq PopSize], TC_i \leftarrow \sum_{j=1}^{n} DM(Indiv_i(j), Indiv_i(j+1)) , \ \ j+1 \equiv 1$$

Step 12: Evaluate the Pollution of each Individual in the Population

$$\forall [1, \leq i \leq PopSize], TP_i \leftarrow \sum_{j=1}^{n} POLL(Indiv_i(j), Indiv_i(j+1)) , \ \ j+1 \equiv 1$$

Step 13: $Gen \leftarrow Gen + 1$//Current generation is completed, increment the Gen for next generation

Step 14: Return Pop

So, the process of inclusion of the city $C2$ adjacent to the city $C1$ in the route is aborted and the condition is verified with the next least normalized value city of $C1$ which is $C3$. The same procedure is repeated until the complete route is generated with n number of cities and the possible route would be $(C1, C3, C2, C4, C6, C5, C7, C1)$. It not guaranteed that the individuals in the population yields optimal solution to the problem for both air pollution and distance, based on the genetic operations the individuals in the populations are improved.

The objective of intelligent routing strategy for VRP in Hybrid Optimal routing is to minimize the air pollution and the distance of the individuals in the population, for that different tradeoff method should be provided.

IIOAB JOURNAL
ISSN: 0976-3104

Algorithm explanation

The algorithm for Hybrid Optimal routing has the following arguments; Pop is the initial population generated using random or heuristic technique, TC is the total cost of each individual in the initial population using equation (1), G is the generation limit for termination of GA and n is the size of the problem instance. Elitism Rate (ER) is the number of high quality / elitist individuals are transferred from the current generation to the next without any modification. This elitism transfer technique avoids the replacement of best fit individuals with poor individuals in the successive generations and also improves the performance of crossover operation, if the parent is selected from the elitist individuals. The total cost and total pollution of each individual in the population is determined and represented as TC_j and TP_j. The total cost convergence rate TCC and the total pollution convergence rate TPC of the individuals in the population is derived through equations (2) and (3). The average of both pollution and distance convergence has been computed (i.e.) $TradeOff\_Con_1$ of the each individual in the population. The ER numbers of individuals having best fitness (i.e.) maximum tradeoff convergence value are hand-picked based on the position and send to consecutive generation.

$$TC_j = \begin{vmatrix} TC_1 \\ TC_2 \\ TC_3 \\ \cdots \\ TC_{PopSize} \end{vmatrix} = \begin{vmatrix} DM(\text{Indiv}_1) \\ DM(\text{Indiv}_2) \\ DM(\text{Indiv}_3) \\ \cdots \\ DM(\text{Indiv}_{PopSize}) \end{vmatrix} = \begin{vmatrix} \text{Indiv}_1 \\ \text{Indiv}_2 \\ \text{Indiv}_3 \\ \cdots \\ \text{Indiv}_{PopSize} \end{vmatrix} = \begin{vmatrix} City_1 & City_i & City_i & \cdots & City_{n-1} & City_1 \\ City_1 & City_i & City_i & \cdots & City_{n-1} & City_1 \\ City_1 & City_i & City_i & \cdots & City_{n-1} & City_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ City_1 & City_i & City_i & \cdots & City_{n-1} & City_1 \end{vmatrix}$$
$$= CPOP_{PopSize \times n} \quad (1)$$

$$TCC \leftarrow (1 - (TC_1 - Opt\_Dist)/Opt\_Dist) * 100 \quad (2)$$

$$TPC \leftarrow (1 - (TP_1 - Opt\_Poll)/Opt\_Poll) * 100 \quad (3)$$

First, two parent solutions P_Indiv1 and P_Indiv2 are chosen randomly from the current population and the first city of the parents is copied as the first city of the off-springs, thus the $Length = 1$. The construction of a complete offspring $Indiv$ of length 'n' using the greedy crossover is explained in the subsequent discussion:
The position of the current cityCur_City of the partially built offspring $Indiv$ in the two selected parents is identified using the following conditions,

$$Pos1 \leftarrow find(P_{indiv1}(Cur\_City)) \quad (4)$$

$$Pos2 \leftarrow find(P_{indiv2}(Cur\_City)) \quad (5)$$

The position of current city in the parents is used to identify the location of left $LLoc$ and right $RLoc$ adjacent cities of Cur_city in the concerned parent solutions and the corresponding location value can be acquired by following the following heuristic:

IF $Pos1 = 1$
$\quad LLoc1 \leftarrow n, RLoc1 \leftarrow Pos1 + 1$
Else IF $Pos1 = n$
$\quad LLoc1 \leftarrow Pos1 - 1, RLoc1 \leftarrow 1$
Else
$\quad LLoc1 \leftarrow Pos1 - 1, RLoc1 \leftarrow Pos1 + 1$
IF $Pos2 = 1$
$\quad LLoc2 \leftarrow n, RLoc2 \leftarrow Pos2 + 1$
Else IF $Pos2 = n$
$\quad LLoc2 \leftarrow Pos2 - 1, RLoc2 \leftarrow 1$
Else
$\quad LLoc2 \leftarrow Pos2 - 1,\ RLoc2 \leftarrow Pos2 + 1$

The location of adjacent cities in the parent solutions are used to find the city with the least distance from the Cur_City is determined,

$$d_1 \leftarrow DM(P\_Indiv1(LLoc1), P\_Indiv1(Pos1)) \quad (6)$$

$$d_2 \leftarrow DM(P\_Indiv1(Pos1), P\_Indiv1(RLoc1)) \quad (7)$$

$$d_3 \leftarrow DM(P\_Indiv1(LLoc2), P\_Indiv1(Pos2)) \quad (8)$$

$$d_4 \leftarrow DM(P\_Indiv1(Pos2), P\_Indiv1(RLoc2)) \quad (9)$$

IIOAB JOURNAL
ISSN: 0976-3104

The location of adjacent cities in the parent solutions are used to find the city with the least air pollution from the Cur_City is determined,

$$p_1 \leftarrow POLL(P\_Indiv1(LLoc1), P\_Indiv1(Pos1)) \quad (10)$$
$$p_2 \leftarrow POLL(P\_Indiv1(Pos1), P\_Indiv1(RLoc1)) \quad (11)$$
$$p_3 \leftarrow POLL(P\_Indiv1(LLoc2), P\_Indiv1(Pos2)) \quad (12)$$
$$p_4 \leftarrow POLL(P\_Indiv1(Pos2), P\_Indiv1(RLoc2)) \quad (13)$$

Normalizing the calculated adjacent cities distance and pollution values using equation (14) and (15).Ω_z Represents the tradeoff values for distance and pollution of each adjacent city that is estimated using equation (16).

$$\forall[1, \leq t \leq 4], D'_t = \frac{D_t}{\sum_{i=1}^{4} D_i} \quad (14)$$
$$\forall[1, \leq t \leq 4], P'_t = \frac{P_t}{\sum_{i=1}^{4} P_i} \quad (15)$$
$$\forall[1, \leq z \leq 4], \Omega_z \leftarrow D'_z + P'_z \quad (16)$$

The least tradeoff value among the four $\Omega_1, \Omega_2, \Omega_3 and\ \Omega_4$ is selected and the city at the corresponding location of the concerned parent is chosen as the next city $Next\_City$. The chosen city $Next\_City$ is verified for two conditions,

Condition 1: The chosen city should not present in the partially built offspring i.e. $Next\_City \notin Indiv$.

Condition 2: The pollution value between the current city Cur_City and the chosen next city $Next\_City$ should be within the maximum pollution limit $POLL_{Cur\_City, Next\_City} < Max\_Poll$.

If the chosen city satisfies both the conditions, it is added as the next city in the offspring$Indiv$ and the length of the offspring is incremented, $Length \leftarrow Length + 1$ otherwise the city with next least distance is chosen and verified. If all the possible cities are checked, next city is added randomly. The same steps are repeated until the length of the offspring $Indiv1$ is n which indicates that the offspring is a feasible solution/route of n cities. The similar procedures are followed to construct the second offspring $Indiv2$. The swap mutation is applied at the resultant offspring's by exchanging the randomly chosen cities equation (17) and (18), within the offspring as,

$$GeneC1 \leftarrow RAND(1, n,) \quad (17)$$
$$GeneC2 \leftarrow RAND(1, n) \quad (18)$$
$$Indiv(GeneC1) \leftarrow Indiv(GeneC2) \quad (19)$$
$$Indiv(GeneC2) \leftarrow Indiv(GeneC1) \quad (20)$$

This stage confirms that the construction of offspring is completed and it is included in the next population and the size of the population is incremented $Size \leftarrow Size + 1$. The generation of next population Pop of individuals is said to be completed if the $Size = PopSize$ and the population generation is repeated for G number of times, then the execution stops. The final population is assessed for the best solution in terms of distance and pollution using equation (21) and (22) respectively.

$$\forall[1, \leq i \leq PopSize], TC_i \leftarrow \sum_{j=1}^{n} DM(Indiv_i(j), Indiv_i(j+1)), \quad j+1 \equiv 1 \quad (21)$$
$$\forall[1, \leq i \leq PopSize], TP_i \leftarrow \sum_{j=1}^{n} POLL(Indiv_i(j), Indiv_i(j+1)), \quad j+1 \equiv 1 \quad (22)$$

METHODS

As discussed in section 3, The Hybrid Optimal based Routing in TSP is based on the tradeoff between the distance and air pollution exploring this problem as a multi objective. The intension is to find the optimal route based on "the total distance of the route" and "the total air pollution of the route". In each of the performance criteria associated with this scenario, the cost refers to the total distance of the solution obtained.

IIOAB JOURNAL
ISSN: 0976-3104

Hybrid optimal routing in VRP

In this scenario of experiments, the intelligent routing in VRP has been performed by optimizing the total distance of the route and the total air pollution of the route. Experimental results for this scenario were analyzed with Random, Nearest Neighbor and ODV based EV population seeding techniques are shown in the [Table 1]. From the [Table 1] following observations can be made:

Observation 1: For all the problem instances, the ODV-EV population seeding technique yields higher convergence rate for the best individual within the population w.r.t air pollution and distance. In best convergence rate or the maximum convergence rate for distance obtained in ODV-EV technique is 98.830% for eil51 and for the air pollution the Maximum of 100% obtained in ODV-EV technique for eil51. The minimum convergence rate for distance and air pollution are 40.407% and 57.213% obtained in random technique for the instance KroA100.

Observation 2: It is observed from the result that the worst convergence rate or the worst individuals in the population of ODV-EV technique showed better performance. The maximum and minimum convergence rate obtained in NN and Random technique are 57.139% and -42.884 for distance. The worst individuals acquired for Air pollution with the maximum rate of 76.864 and the minimum rate of -53.138.

Observation 3: Performance analyses in distance based on the error rate reveal that the ODV-EV technique performs outstandingly and has maximum of 19.234% for the instance Swiss42 where as NN and random techniques have maximum of 40.412% and 59.593% respectively for the instance KroA100. The minimum and maximum worst error rate in terms of air pollution obtain from that worst individuals are 23.136 % in ODV-EV for the instance KroA100 and 153.138% in NN technique for the instance uysses22.

Observation 4: The average convergence is working better in ODV –EV technique, Average Convergence for Hybrid Optimal routing (Pollution) is less in small cities, and it increases gradually when we are moving towards the large instances. The average convergence of NN technique is less than the random technique. As it is the evident from [Table 1], both the minimum and maximum average convergences for pollution is obtained in ODV-EV technique ranges from -18.298 to 84.636 in uysses22 and eil51 respectively.

Observation 5: The Convergence diversity of distance as well as pollution values of all the instances is better in ODV-EV technique. The minimum and maximum values of convergence rate w.r.t distance are acquired in random technique are 38.684 and 91.535. The convergence diversity w.r.t pollution in ODV-EV technique is obtained the minimum value is 18.545 and the maximum value is 150.29.

Observation 6: It is observed from the [Table 1] that, the computational time varies for each problem instances. Based on the Problem Instance, ODV-EV and NN techniques showed a gradual increase. The performance of random technique is irregular and it showed unexpected changes in the computation time for the instance eil76. The ODV-EV technique performs well and has less computational time for the entire instance except eil76.

RESULTS

In this section we have discussed each and every performance factors of the proposed system and also from the result we have identified ODV-EV seeding technique performs better than other seeding techniques. The rest of the section is evident the performance of the proposed system.

Best Convergence rate w.r.t distance: The best convergences of distance in ODV-EV technique of are performing well when compared to the best convergence of distance in other techniques. The NN technique is performing better than the random technique in many instances as showed in the [Fig.2]. The performance of random technique for each instance is uneven, gradually decreased, then increased and then decreased.

Best Convergence rate w.r.t pollution: From the Graph it is analyzed that, the best convergence of pollution in NN and Random techniques have showed lower performance, when compare to the best convergences of pollution in ODV-EV technique. Most of the time the performance of random technique is superior to the NN technique in many instances is showed in the [Fig.8]. Moving towards the higher instances the convergence - rate is gradually decrease in ODV-EV and NN technique, in case of random technique sudden decrease in best convergence rate.

Best Error rate w.r.t distance: the best convergence rate is high in ODV-EV; obviously the best error rate is less in ODV-EV technique. Because, the pollution in the path between current city and the next minimum distance city is high. It will move to the next minimum distance city. So the convergence rate is high and the error rate is less. The random technique has higher error rate than the NN technique in many instances is showed in the [Fig.3], shows that the performance of NN technique is better than the random technique.

IIOAB JOURNAL
ISSN: 0976-3104

Best Error rate w.r.t pollution: The [Fig.9] showed that the Best convergence rate is higher for every instance, which clearly states that the best error rate of ODV-EV is lower than other techniques. This openly implies the quality of the individuals in the population of the desire technique. Based on comparison and analysis of NN and Random technique, random has satisfactory performance projected in [Fig. 9].

Convergence diversity w.r.t distance: The convergence diversity is an aspect that illuminates the distribution of good and bad quality individuals among the population. It plays a vital role to increase the chance of evolving optimal solutions and to avoid premature convergence. [Fig.7] shows the convergence diversity of the optimal distance based routing scenario using different population seeding techniques for the problem instances. From the [Fig.7], it is understood that the ODV-EV technique has lesser convergence diversity w.r.t. other population seeding techniques which shows that the quality of individuals is improved as a population rather than the single individual. For most of the instances, random and NN techniques have nearly equal convergence diversity.

Convergence diversity w.r.t. Air Pollution: The convergence diversity of the air pollution based optimal routing scenario using different population seeding techniques for the problem instances in shown in the [Fig.13]. From the [Fig.13], it is observed that the convergence diversity of the instances decreases with increase in the problem size despite the population technique applied.

Average convergence w.r.t distance: The Average convergence of a population is used to measure the quality of the population generated by finding the average of fitness of individuals in the population as shows in [Fig.6] the average convergence rate for hybrid optimal routing (distance) using different population seeding techniques for the problem instances. From the [Fig.6], it can be observed that every population seeding technique yields better average convergence rate for some of the large size problem instances than the small size instances. For most of the instances, the ODV-EV technique outperforms other population initialization techniques and random performs worst for the larger size instances. For the instance bays29, performance of random, NN and ODV-EV techniques are very poor; this is possibly because of the peculiarity of the instance with small size and large distance based fitness value.

Average convergence w.r.t pollution: [Fig.12] shows the average convergence rate for air pollution based optimal routing using different population seeding techniques for the problem instances. From the [Fig.12], it can be understood that average convergence rate increases with increase in the size of the problem instances regardless of the population technique used. In the case of average convergence rate, all the population seeding techniques perform nearly equal though ODV-EV technique yields marginally better result than other techniques.

Average error rate w.r.t distance: The average error rate is working better in ODV –EV technique, compare to other techniques. Average error rate of NN technique is less than the random technique is showed in [Fig.14]. The performance of random technique is unpredictable; it shows huge variation for each problem instance, this evidently indicates that the quality of the individuals in the population is less.
Average error rate w.r.t pollution: the Average error rate w.r.t pollution, the ODV-EV technique shows high values in smaller instances and performance is increases as increase in problem instance. The [Fig.15] exposed, that the average convergence of NN technique showed a reasonable output for all the instances. The analysis shows that performance of NN technique in terms of average convergence is better than other techniques.
[
Worst Convergence rate w.r.t distance: [Fig.4] shows the worst convergence rate for Worst Convergence Rate for Hybrid Optimal routing (Distance) using different population seeding techniques for the problem instances. From the [Fig.4], it can be observed that ODV-EV technique yields better results than the NN and random technique. For the instance eil51, the random technique outperforms than the other techniques.

Worst Convergence rate w.r.t Pollution: The worst convergence rate of distance and pollution in ODV- EV technique is good, than the other two techniques. [Fig. 10] shows the worst convergence rate for optimal pollution based routing using different population seeding techniques for the problem instances. From the [Fig.10], it is observed that every population seeding technique yields better, worst convergence rate for some of the large size problem instances than the small size instances.

IIOAB JOURNAL
ISSN: 0976-3104

Instance	Seeding Technique	Model	Optimal Solution	Computation Time	Quality of the Solution			Convergence Rate (%)		Error Rate (%)		Convergence Diversity	Average Convergence
					Best	Worst	Average	Best	Worst	Best	Worst		
		Pollution	2.5596		2.849	6.120	5.092	88.686	-39.100	11.314	139.100	127.786	1.044
	EV	distance	74.1087	11.230	82.465	124.907	109.448	88.725	31.454	11.275	68.546	57.270	52.315
		Pollution	2.5596		2.915	6.418	5.070	86.134	-50.737	13.866	150.737	136.871	1.923
	NN	distance	74.1087	11.370	84.086	130.483	113.127	86.537	23.930	13.463	76.070	62.606	47.350
		Pollution	2.5596		2.887	6.378	4.918	87.211	-49.187	12.789	149.187	136.399	7.863
uysses16	Random	distance	74.1087	11.270	83.914	135.927	110.031	86.768	16.585	13.232	83.415	70.184	51.528
		Pollution	3.194		3.462	8.002	6.972	91.603	-50.537	8.397	150.537	142.140	-18.298
	EV	distance	75.6615	16.700	80.566	129.462	129.719	93.518	28.893	6.482	71.107	64.624	28.553
		Pollution	3.194		3.285	8.085	6.529	97.152	-53.138	2.848	153.138	150.290	-4.429
	NN	distance	75.6615	17.190	93.653	155.851	131.725	76.221	-5.984	23.779	105.984	82.205	25.902
		Pollution	3.194		2.976	7.872	6.136	83.762	-46.474	16.238	146.474	130.236	7.878
uysses22	Random	distance	75.6615	16.850	95.554	148.302	132.271	73.708	3.993	26.292	96.007	69.715	25.181
		Pollution	5.1614		5.609	9.234	9.429	91.330	21.095	8.670	78.905	70.235	17.317
	EV	distance	2020	21.860	2186.000	3842.800	4024.220	91.782	9.762	8.218	90.238	82.020	0.781
		Pollution	5.1614		6.511	10.454	8.932	73.861	-2.544	26.139	102.544	76.406	26.955
	NN	distance	2020	23.830	2800.000	4622.600	3990.376	61.386	-28.842	38.614	128.842	90.228	2.457
		Pollution	5.1614		6.186	10.773	9.331	80.153	-8.724	19.847	108.724	88.877	19.209
bays29	Random	distance	2020	23.510	2946.600	4795.600	4149.330	54.129	-37.406	45.871	137.406	91.535	-5.412
		Pollution	6.2613		7.466	8.762	9.561	80.766	60.061	19.234	39.939	20.705	47.297
	EV	distance	1273	23.550	1330.600	1918.800	1642.133	95.475	49.269	4.525	50.731	46.206	71.003
		Pollution	6.2613		8.301	9.925	9.219	67.416	41.489	32.584	58.511	25.928	52.766
	NN	distance	1273	22.730	1474.400	2134.400	1807.974	84.179	32.333	15.821	67.667	51.846	57.975
		Pollution	6.2613		8.120	9.785	9.028	70.314	43.727	29.686	56.273	26.587	55.808
swiss42	Random	distance	1273	23.040	1486.400	2050.400	1777.124	83.236	38.932	16.764	61.068	44.305	60.399
		Pollution	7.6588		7.659	10.669	8.835	100.000	60.697	0.000	39.303	39.303	84.636
	EV	distance	426	28.290	430.983	683.409	516.770	98.830	39.575	1.170	60.425	59.255	78.693
		Pollution	7.6588		8.257	11.268	10.016	92.183	52.872	7.817	47.128	39.311	69.227
	NN	distance	426	28.850	471.989	725.836	606.292	89.205	29.616	10.795	70.384	59.589	57.678
		Pollution	7.6588		7.974	10.073	9.285	95.879	68.484	4.121	31.516	27.395	78.768
EIL51	Random	distance	426	29.550	443.793	608.587	543.201	95.823	57.139	4.177	42.861	38.684	72.488
		Pollution	11.3454		11.554	15.132	13.993	98.163	66.620	1.837	33.380	31.543	76.667
	EV	distance	538	48.020	551.174	850.545	761.941	97.551	41.906	2.449	58.094	55.645	58.375
		Pollution	11.3454		13.157	16.341	15.162	84.036	55.971	15.964	44.029	28.064	66.356
	NN	distance	538	46.190	632.503	901.407	798.463	82.434	32.452	17.566	67.548	49.982	51.587
		Pollution	11.3454		13.199	16.510	15.207	83.658	54.477	16.342	45.523	29.181	65.959
eil76	Random	distance	538	30.470	638.987	950.171	805.874	81.229	23.388	18.771	76.612	57.841	50.209
		Pollution	14.5057		15.172	17.862	17.096	95.408	76.864	4.592	23.136	18.545	82.143
	EV	distance	21285	55.690	21918.398	33873.150	30479.214	97.024	40.859	2.976	59.141	56.165	56.804
		Pollution	14.5057		19.222	22.244	21.007	67.490	46.652	32.510	53.348	20.838	55.184
	NN	distance	21285	60.120	29886.677	46054.165	40052.584	59.588	-16.369	40.412	116.369	75.957	11.827
		Pollution	14.5057		20.712	23.666	22.315	57.213	36.849	42.787	63.151	20.364	46.162
kroA100	Random	distance	21285	64.220	33969.432	51697.878	43604.118	40.407	-42.884	59.593	142.884	83.291	-4.858

Table 1: Result analysis of Hybrid optimal based routing

Worst Error rate w.r.t distance: [Fig.5] indicate that the ODV-EV technique performing better than the NN and random technique and it got lower values for all the instances. Although the performance of ODV-EV technique showed good result in worst error rate, the random technique showed a least value for the instance eil51. For most of the instances the worst error rate of NN technique is lesser than the random technique, infers the NN performance is better than the random technique.

Worst Error rate w.r.t Pollution: As shown in the figure [Fig.11], it is clearly perceptible that the worst convergence rate of pollution in ODV- EV technique is virtuous, apart from the other two techniques. It has been observed that, the performance of worst convergence rate in terms of pollution in NN and random technique are inversely proportional to the performance of worst convergence rate in terms of distance. In pollution, the performance of random technique is superior to the NN technique.

Computational Time: [Fig.16] significantly proves that the computation time increases based on the problem instances, each technique has its own computation time for every problem instances. In terms of computation time, it is obvious that the random technique showed good result in classical TSP or any other problem. In this

IIOAB JOURNAL
ISSN: 0976-3104

case, each technique should validate the pollution between the corresponding cities before adding the next city. Hence, the computation time of each technique for different instances has slight changes. Furthermore, analyzed from the [Fig.16] the random technique has showed an abnormal change for the instance eil51, except that the ODV-EV technique shows good performance.

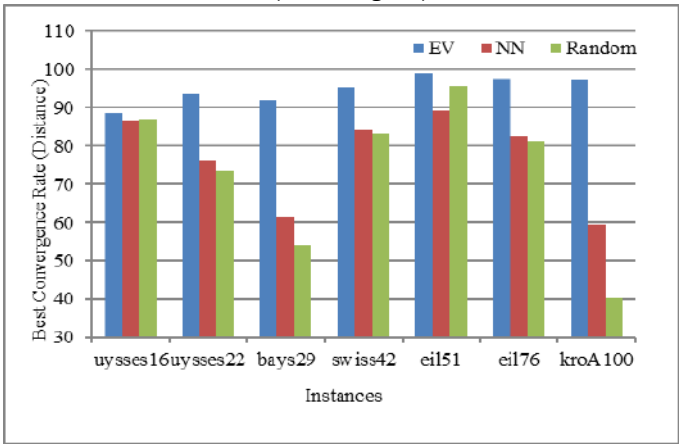


Fig.2: Best convergence rate for hybrid optimal routing (Distance).

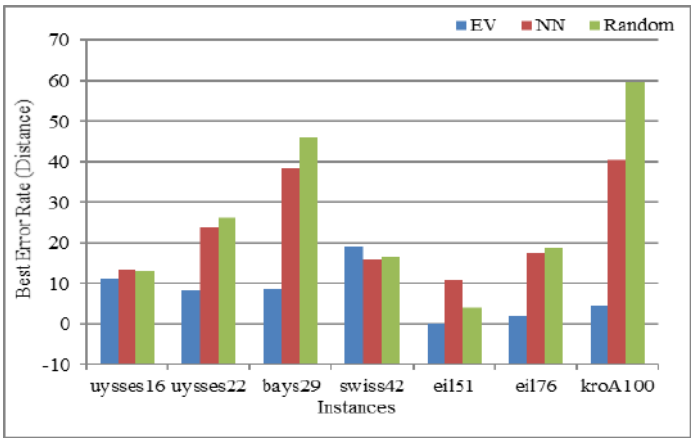


Fig.3: Best error rate for hybrid optimal routing (Distance).

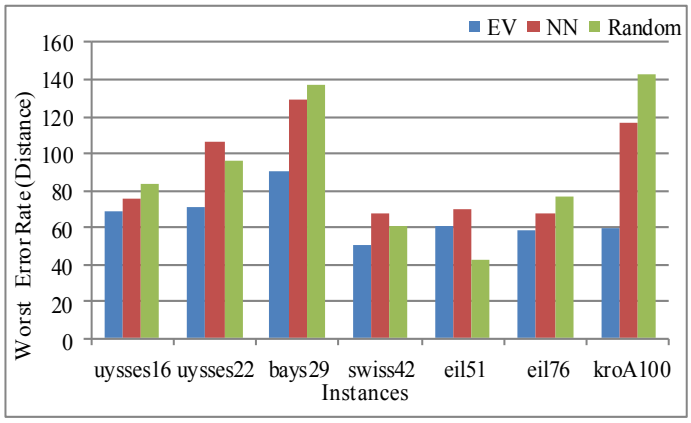


Fig. 4: Worst Convergence Rate for Hybrid Optimal routing (Distance).

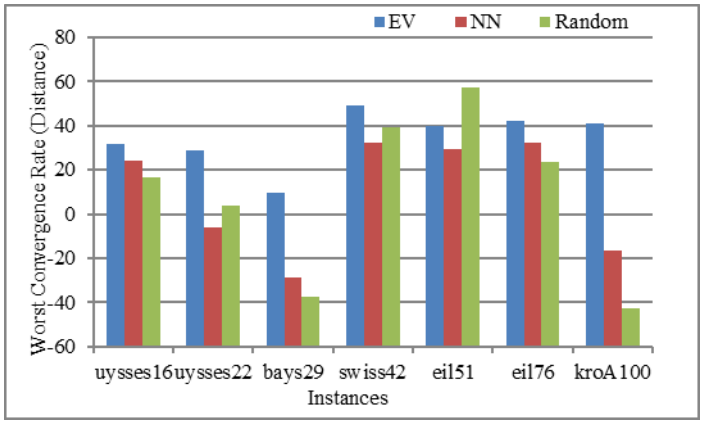


Fig. 5: Worst error rate for hybrid optimal routing (Distance).

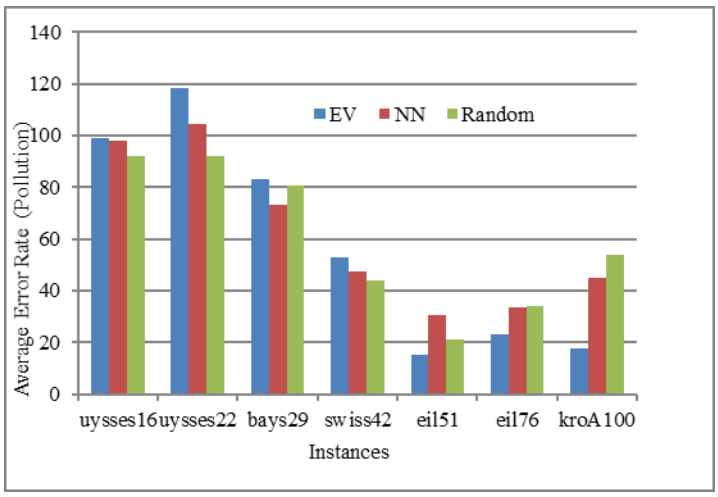


Fig. 14: Average error rate for hybrid optimal routing (Distance).

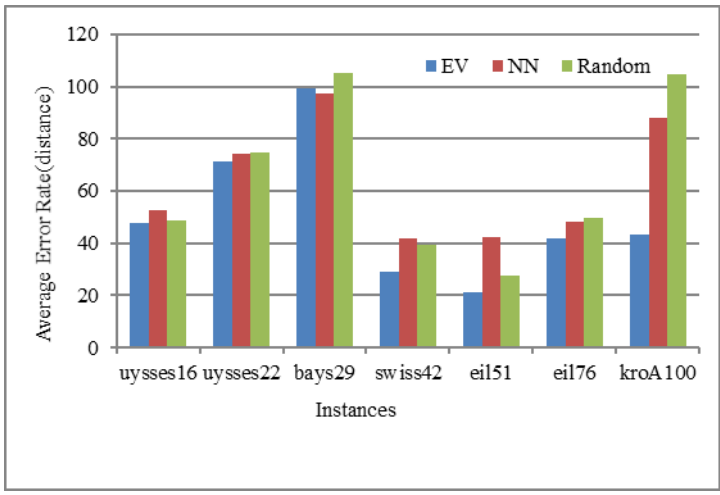


Fig. 15: Average error rate for hybrid optimal routing (Pollution).

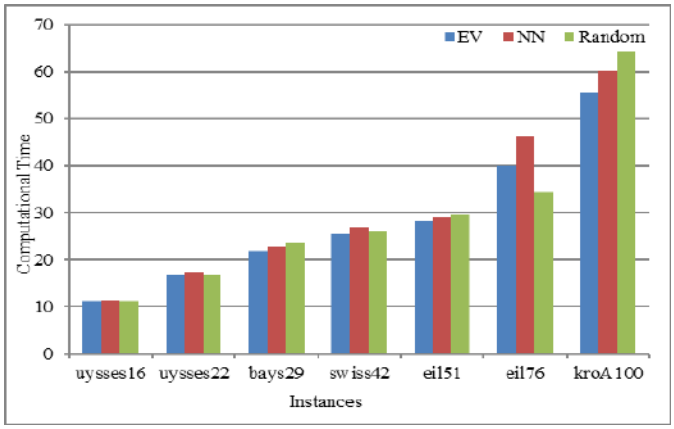


Fig.16: Computational Time for Hybrid Optimal routing.

CONCLUSION

In summary, we proposed and investigated hybrid optimal based routing with different GA initialization techniques like Random, NN and ODV-EV techniques respectively for finding the best optimal (pollution free as well as minimum distance) route for transportation system. We have analyzed our algorithm with standard TSP bench marks and we created the corresponding air pollution matrix, for the instances ulysses16, ulysses22, bays29, att48, eil56, eil76 and kroA100. In the result analysis, we have analyzed our algorithm with different validation criteria's like Best convergence rate, worst convergence rate, average convergence rate, Best error rate, worst error rate and convergence diversity. The algorithm performed well in the ODV-EV technique for all the TSP instances. Next The NN technique is performing better in many instances than the random technique. The ODV-EV technique for optimal distance based routing yielding the best distance convergence of 95.560 % in the instance eil51 and for the optimal pollution based routing yields the best pollution convergence 98.29% in the instance eil76. The hybrid optimal based routing algorithms best convergence rate of distance and pollution are 98.830 % and 100 % in the instance eil51. Since we are calculating the distance from the display coordinates, we are not getting 100%

IIOAB JOURNAL
ISSN: 0976-3104

best convergence rate. From results we analyzed that the ODV-EV technique is performing well. To improve the smartness of ITS we are focusing on enacting this proposed green computing VRP model with VANET for providing next generation ITS.

CONFLICT OF INTEREST

We, Dr. M. Shanmugam* and Dr. J. Amudhavel authors of the manuscript titled "REVENANT OF THE ECOSYSTEM: AN ENVIRONMENTAL BASED GREEN COMPUTING MODELS FOR VEHICULAR ROUTING PROBLEMS USING GENETIC ALGORITHM OPTIMIZATION APPROACH" declaring that there is no conflict of interest regarding the publication of this paper in Institute of Integrative Omics and Applied Biotechnology-IIOAB.

ACKNOWLEDGEMENTS

This work is not part of any organization or institution.

FINANCIAL DISCLOSURE

This work doesn't get any financial assistance.

REFERENCES

[1] Shanmugam M, Saleem Basha MS, Dhavachelvan P, Baskaran R. [2013] Performance Assessment over Heuristic Population Seeding Techniques of Genetic Algorithm: Benchmark Analyses on Traveling Salesman Problems, International Journal of Applied Engineering Research, Research India Publications, 10(8):1171-1183.
[2] Ling-NingXing, Ying-Wu Chen, Ke-Wei Yang, FengHou, Xue-Shi Shen, Huai-Ping Cai. [2008] A hybrid approach combining an improved genetic algorithm and optimization strategies for the asymmetric traveling salesman problem, Engineering Applications of Artificial Intelligence 21:1370–1380.
[3] Albayrak M, Allahverdi N.[2011] Development a new mutation operator to solve the traveling salesman problem by aid of genetic algorithms, Expert Systems with Applications 38:1313–1320.
[4] Helsgaun K.[2000] An effective implementation of the Lin–Kernighan traveling salesman heuristic, European Journal of Operational Research 126(1):106–130.
[5] Marinakis Y, Migdalas A, Pardalos PM.[2005] Expanding neighborhood GRASP for the traveling salesman problem. Computational Optimization and Applications 32: 231–257.
[6] Milthers NPM. [2009] Solving VRP using Voronoi Diagrams and Adaptive Large Neighborhood Search. Master's thesis, University of Copenhagen, Denmark.
[7] Vose MD.[1998] The simple genetic algorithm: foundations and theory." Cambridge, MA: MIT Press, Book
[8] Subramanian A, Penna PHV, Uchoa E, Ochi LS.[2012] A hybrid algorithm for the heterogeneous fleet vehicle routing problem. European Journal of Operational Research 221 :285–295.
[9] Anand Subramanian., Eduardo Uchoa .[2010] Luiz Satoru Ochi., A hybrid algorithm for a class of vehicle routing problems, Computers & Operations Research 40:2519–2531.
[10] Martin Josef Geiger. Fast Approximation Heuristics for Multi-Objective Vehicle Routing Problems. In: Di Chio C et al. (Eds.): Evo Applications, Part II, LNCS 6025
[11] Shubhra SR, Sanghamitra B, Sankar KP.[2007] Genetic operators for combinatorial optimization in TSP and microarray gene ordering, Journal of Applied Intelligence 26: 183-195.
[12] Yannis Marinakis, Magdalene Marinakib, Georgios Dounias.[2010] A hybrid particle swarm optimization algorithm for the vehicle routing problem, Engineering Applications of Artificial Intelligence 23 :463–472.
[13] Yannis Marinakis, Georgia-Roumbini Iordanidou., Magdalene Marinaki. [2013]Particle Swarm Optimization for the Vehicle Routing Problem with Stochastic Demands, Applied Soft Computing 13: 1693–1704.
[14] Kennedy J, Eberhart R.[1995] Particle swarm optimization. Proceedings of 1995 IEEE International Conference on Neural Networks 4: 1942–1948.
[15] Ropke S, Pisinger D.[2006] An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. Transportation Science 40 (4):455–472.
[16] Mostepha R, Khouadjia, Briseida Sarasola, Enrique Alba, Laetitia Jourdan El, GhazaliTalbi A.[2012] comparative study between dynamic adapted PSO and VNS for the vehicle routing problem with dynamic requests, Applied Soft Computing 12:1426–1439.
[17] Gorges-Schleuter M.[1997] Asparagos96 and the traveling salesman problem, Proceedings of IEEE International Conference on Evolutionary Computation, Indianapolis, 171–174.
[18] Merz P, Freisleben B.[1997] Genetic local search for the TSP: new results. In: Proceedings of the IEEE International Conference on Evolutionary Computation. IEEE Press, Indianapolis, 159–164.
[19] Ross BJ.[1999] A Lamarckian evolution strategy for genetic algorithms, Practical Handbook of Genetic Algorithms: Complex Coding Systems, 3: 1–16.
[20] Norstad I.[2011] Agerholt, Laporte, G, Tramp ship routing and scheduling with speed optimization. Transportation Research Part C: Emerging Technologies & Science 5: 853–865.
[21] Winter G, Périaux J, Galaán M, Cuesta P.[1995] Genetic algorithms in engineering and computer science, New York: Wiley
[22] Jozefowiez N, Semet F, Talbi EG.[2008] Multi-objective vehicle routing problems. European Journal of Operational Research 189(2):293–309.
[23] Park YB, Koelling CP.[1989] An interactive computerized algorithm for multi-criteria vehicle routing problems. Computers & Industrial Engineering 16(4):477–490.
[24] Solomon MM.[1987] Algorithms for the vehicle routing and scheduling problems with time windows constraints. Operations Research 35(2):254–265.
[25] Marinakis Y, Migdalas A, Pardalos PM.[2005] A hybrid genetic–GRASP algorithm using Lagrangian relaxation for the traveling salesman problem. Journal of Combinatorial Optimization 10: 311–326.
[26] Victer Paul, Ramalingam P, Baskaran A, et al. [2013] Performance Analyses on Population Seeding Techniques for Genetic Algorithms, International Journal of Engineering and Technology 5 (3):2993-3000.

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A SIMPLE LOSSLESS COMPRESSION ALGORITHM IN WIRELESS SENSOR NETWORKS: AN APPLICATION OF SEISMIC DATA

J. Uthayakumar[1*], T. Vengattaraman[1], J. Amudhavel[2]

[1] Department of Computer Science, Pondicherry University, Puducherry, INDIA
[2]Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA

ABSTRACT

Background: Seismic hazard is one of the natural hazards which are very difficult to predict and detect. Various seismic and seismo-acoustic monitoring systems are developed to observe the changes in rock mass processes. Some of the recent seismic dataset exceed 10Tbytes and there is a fact that present seismic studies planned with an amount of 120Tbytes. The size of the seismic data is a challenge to store, process or transmit seismic data for further investigation. ***Methods:*** Data compression (DC) techniques are commonly used to reduce the amount of data being stored or transmitted. As seismic monitoring system is a critical application, the loss of data quality is not tolerable. So, lossless data compression techniques are highly preferable. In this paper, Burrows Wheeler Transform (BWT) is used to compress the seismic data effectively. BWT is a block sorting compression which rearranges the character string into runs of identical characters. ***Results:*** *Extensive experiments were performed using real world seismic bump dataset. To ensure the effectiveness of lossless BWT coding, it is compared with 4 well-known compression algorithms namely Lempel Ziv Markov Algorithm (LZMA), Huffman coding, Lempel-Ziv-Welch (LZW) coding and Arithmetic coding (AC).* ***Conclusions:*** *By comparing the compression performance of BWT coding with existing methods, BWT achieves significantly better compression with a compression ratio of 0.167 at the bit rate of 1.342 bpc.*

KEY WORDS
Data Compression
BWT coding
Seismic data
Wireless Sensor Network

Received: 3 June 2017
Accepted: 28 July 2017
Published: 18 Sept 2017

***Corresponding Author**
Email:
uthayresearchscholar@gmail.com
Tel.: +91 9677583754

INTRODUCTION

The unexpected emission of energy from the earth crust results to earthquake. Because of the energy, seismic waves arise and the waves generate a distress effect on earth. There are various issues that influence earthquake. These factors enhance the complexity and it leads to identify location, intensity and time of the earthquake. Several studies use several information to predict earthquake like earth's crust form change, slope change, radon gas changes in well and springs, elastic variable wave velocities, groundwater level variation and seismic pulses. Few researches used seismic bumps to carry out this process. [1] developed a system to predict earthquake by investigating seismic bump data. They used k nearest neighbor algorithm from classification and obtained an accuracy of 94.11%. [2] introduced an intelligent system to predict earthquake by investigating seismic bump data with a classification accuracy of 91%. [3] performed an experiment using neuro-fuzzy system algorithm in seismometer data and they achieved an accuracy of 82%. Generally, the seismic sensors are deployed in the surface which measures the seismic signals and the signals are processed to create the picture of the subsurface. Some advanced marine seismic dataset are larger in size and it exceeds 10 Tbytes. In some cases, the volume of seismic data may exceed 120Tbytes. Since, there is a need to compress this massive amount of seismic data. Data transmission is the most energy consuming task due to the nature of strong temporal correlation in the sensed data. DC is considered as a useful approach to eliminate the redundancy in the sensed data. In WSN, the sensor node runs the compression algorithm and the compressed data will be forwarded to Base Station. The easier way to handle huge amount of data is to compress them [3]. It is applicable to compress text, images, audio or video. The data can be alphanumeric characters in a text document or numbers which represent the samples in an audio or image waveforms or series of numbers created by some processes, etc. DC can also be termed as an efficient way of representing the data in its compact manner. The compact form can be achieved by the recognition and utilization of patterns exists in the data. The evolution of DC starts with the Morse code, developed by Samuel Morse in the year 1838 [4]. Morse code is used in telegraph to compress letters. It uses smaller sequences to represent letters which occurs more frequently and thereby minimizing the message size and transmission time. This basic principle is employed in Huffman coding [5]. Nowadays, DC techniques are very essential in most of the real time applications like medical imaging, satellite imaging, Wireless Sensor Networks (WSN), etc. DC became important for the utilization of available resources effectively. Without DC, it is very difficult and sometimes impossible to store or communicate huge amount of data files.

Basically, DC techniques falls under two categories: lossless compression and lossy compression. As the name implies, lossless compression refers that no loss of information i.e. the reconstructed data is exactly same as original data. It is used in situations where the loss of information is unacceptable. Example: Text, medical imaging, satellite imaging, etc. Lossless compression is generally used to compress text. As loss of information is intolerable in text compression, compressing text using lossless compression is mandatory. It achieves the compression ratio of 2:1 to 8:1 [6]. In some situations, the reconstructed data need not to be perfectly matched with the original data and the approximation of original data is also acceptable. In those situations, lossy compression is used which involves with some loss of information within the acceptable level. This leads to higher compression ratio when compared to lossless compression. Example: image, audio and video. Lossy compression is commonly used to compress images as the loss of image quality is tolerable. The main principle behind image compression is the correlation of pixels. The neighboring pixels of an image are high correlated. These redundant details are eliminated by computing a less correlated representation of the image. Lossy compression achieves the compression ratio of 100:1

to 200:1 based on the data to be compressed. It achieves higher compression ratio at a cost of loss in quality.

Only few researches have been made on the compression of seismic dataset. As the lossy compression techniques involve loss of quality, it cannot be used to compress seismic data which is intolerant to errors. It leads to the development of lossless compression of seismic data in the real world. Once the seismic data is compressed, it can be useful in several ways. In WSN, the sensor node executes the compression algorithm to compress the seismic data. Then, the compressed data will be transmitted to BS for further investigation. It can be easily shared; it consumes less storage with low management cost. The absence of effective lossless compression algorithm for seismic data motivated us to perform this work.

Contribution of the paper

The contribution of the paper is summarized as follows: (i) A lossless BWT compression algorithm is used to compress seismic bump dataset in WSN (ii) Two seismic dataset from NREL (Eastern and Western Dataset) is used, and (iii) BWT results are compared with 4 well-known compression algorithms namely Huffman coding, AC, LZW and LZMA interms of Compression Ratio (CR), Compression Factor (CF) and bits per character (bpc).

Organization of the paper

The rest of the paper is organized as follows: Section 2 explains the different types of classical DC techniques in detail. Section 3 presents the BWT compression algorithm for seismic dataset in WSN. Section 4 discusses the compression results obtained for the applied dataset. Section 5 concludes with the highlighted contributions, future work, and recommendations.

RELATED WORK

DC is an efficient way to reduce the amount of data being stored or transmitted. The popular coding methods are Huffman coding, Arithmetic coding, Lempel Ziv coding, Burrows-Wheeler transform (BWT), RLE, Scalar and vector quantization.

Huffman coding [7] is the most popular coding technique which effectively compresses data in almost all file formats. It is a type of optimal prefix code which is widely employed in lossless DC. It is based on two observations: (1) In an optimum code, the frequent occurrence of symbols is mapped with shorter code words when compared to symbols that appear less frequently. (2) In an optimum code, the least frequent occurrence of two symbols will have the same length. The basic idea is to allow variable length codes to input characters depending upon the frequency of occurrence. The output is the variable length code table for coding a source symbol. It is uniquely decodable and it consists of two components: Constructing Huffman tree from input sequence and traversing the tree to assign codes to characters. Huffman coding is still popular because of its simpler implementation, faster compression and lack of patent coverage. It is commonly used in text compression.

AC [8] is an another important coding technique to generate variable length codes. It is superior to Huffman coding in various aspects. It is highly useful in situations where the source contains small alphabets with skewed probabilities. When a string is encoded using arithmetic coding, frequent occurring symbols are coded with lesser bits than rarely occurring symbols. It converts the input data into a floating point number in the range of 0 and 1. The algorithm is implemented by separating 0 to 1 into segments and the length of each segment is based on the probability of each symbol. Then the output data is identified in the respective segments based on the symbol. It is not easier to implement when compared to other methods. There are two versions of arithmetic coding namely Adaptive Arithmetic Coding and Binary Arithmetic Coding. A benefit of arithmetic coding than Huffman coding is the capability to segregate the modeling and coding features of the compression approach. It is used in image, audio and video compression.

Dictionary based coding approaches find useful in situations where the original data to be compressed involves repeated patterns. It maintains a dictionary of frequently occurring patterns. When the pattern comes in the input sequence, they are coded with an index to the dictionary. When the pattern is not available in the dictionary, it is coded with any less efficient approaches. The Lempel-Ziv algorithm (LZ) is a dictionary-based coding algorithm commonly used in lossless file compression. This is widely used because of its adaptability to various file formats. It looks for frequently occurring patterns and replaces them by a single symbol. It maintains a dictionary of these patterns and the length of the dictionary is set to a particular value. This method is much effective for larger files and less effective for smaller files. For smaller files, the length of the dictionary will be larger than the original file. The two main versions of LZ were developed by Ziv and Lempel in two individual papers in 1977 and 1978, and they are named as LZ77 [9] and LZ78 [10]. These algorithms vary significantly in means of searching and finding matches. The LZ77 algorithm basically uses sliding window concept and searches for matches in a window within a predetermined distance back from the present position. Gzip, ZIP, and V.42bits use LZ77. The LZ78 algorithm follows a more conservative approach of appending strings to the dictionary.

IIOAB JOURNAL
ISSN: 0976-3104

LZW is an enhanced version of LZ77 and LZ78 which is developed by Terry Welch in 1984 [11]. The encoder constructs an adaptive dictionary to characterize the variable-length strings with no prior knowledge of the input. The decoder also constructs the similar dictionary as encoder based on the received code dynamically. The frequent occurrence of some symbols will be high in text data. The encoder saves these symbols and maps it to one code. Typically, an LZW code is 12-bits length (4096 codes). The starting 256 (0-255) entries represent ASCII codes, to indicate individual character. The remaining 3840 (256-4095) codes are defined by an encoder to indicate variable-length strings. UNIX compress, GIF images, PNG images and others file formats use LZW coding.

A simpler form of lossless DC coding technique is Run Length Encoding (RLE). It represents the sequence of symbols as runs and others are termed as non-runs. The run consists of two parts: data value and count instead of original run. It is effective for data with high redundancy. For example, "ABCABBBBBC" is the input sequence, then the starting 4 letters are assumed as non-run with length 4, and the next 4 letters are assumed as a run with length 5 while B is repeated 5 times. RLE identifies the run of input file, saves the symbol and length of each run. These runs are used to compress the input file and non-runs are kept uncompressed. It is not effective in cases where the redundancy is low and leads to increase in original file size. It is ineffective for less redundant data and leads to increase in compressed file size greater than original file size. Input: YYBCCECCDESSEEEEER and its Output: 2Y1B2C1E2C1D1E2S5E1R. Though RLE is simpler to implement and faster, it fails to achieve better compression when compared to other algorithms. It is used in line drawing, animation, graphic icons, fax and palette-based images like textures.

Quantization is an easiest way to represent larger set of values to a smaller set. The quantizer input can be either scalar or vector. When the input is scalar, then it is called scalar quantizer. When the input is vector, then it is called vector quantizer. Scalar quantization is the simpler and commonly used lossy compression. It is a process of mapping an input value x to a number of output values. Vector quantization (VQ) is a different type of quantization, which is typically implemented by choosing a set of representatives from the input space, and then mapping all other points in the space to the closest representative. Initially, the source output is grouped into blocks or vectors. These vectors are used as the input to quantizer. In both sides of encoder and decoder, a codebook (set of L-dimensional vectors) is available. The vectors in the codebook are termed as code-vectors and each code vector is assigned to an index value. The elements of this code vector are the quantized values of the source output. In order to report the decoder about which code vector was found to be the closest to the input vector, the index value of the code-vector is transmitted. As the decoder also has same codebook, the code vector can be retrieved from its binary index.

Lossless compression techniques are widely used in text compression as loss of information is not tolerable while compressing text. A novel data compression technique is proposed to compress general data using a logical truth [Table 1]. Here, the two bits of the data are represented by only one bit. Using this method, the data bits can be represented by its half bits only. i.e. 128 bits can be represented by 64 bits; 64 bits can be represented by 32 bits; 1GB data can be represented by 512MB and so on. The logic truth table with two inputs and four combinations are given below.

IF A=0 and B=0 then Z=0
IF A=0 and B=1 then Z=Ô
IF A=1 and B=1 then Z=1
IF A=1 and B=0 then Z=î

First, input data is tested to identify whether it is odd or even. If it is odd, extra bit is added at the last (0 is added when the last bit is 0 and vice versa). Then, the data is split into two bits and truth [Table 1] is applied to compress the data. Traditionally, there are two levels to represent digital data but proposed method uses four levels. The proposed method compresses the 20 bit data into 10 bits. This method uses Compression Ratio and Compression Factor as performance metrics. The author fails to compare the proposed method with existing method. It can be employed in wired and wireless scenarios.

Although some of the common compression techniques can be applicable to some preferred applications, the concentration has been on the technique rather than on the application. However, there are certain techniques for which it is impossible to separate the technique from the application. This is because that several techniques depend on the properties or characteristics of the application. Therefore, some compression techniques were developed with focus on particular applications.

A new database compression method is proposed to compress large databases using association rule mining. Association rule mining from a database can identify frequent item sets in the database relations. The extracted association rules represent the repeated item sets in the database and they are compressed to decrease the file size. There are five steps in database compression: i. An effective association rule mining method using a tree called Class Inheritance Tree (CIT) is employed, ii. Collection of compression rules from strong association rules, iii. Compressed space for each compression rule is calculated, iv. A heuristic method is employed to resolve the conflicts of compression rules and v. Compression ratio is computed.

To reduce the Inter Track Interference (ITI) read latency of Shingled Magnetic Recording, two lossless compression mechanisms are used. It is a known fact that most of the files stored in disks are lossless

compressible. When a sector of user data can be losslessly compressed to a particular level, free storage space will become available to store ECC redundancy. This extra space can be used for storing a stronger ECC than normal ECC for the current sector. This leads to decrease the probability of reading one or multiple neighboring tracks for explicit ITI compensation. Lempel-Ziv-Stores-Symanski (LZSS) algorithm is used for compressing several file types stored in hard disks. This method uses intra-sector data compression method to minimize the read latency is shingled recording. Here, lossless data compression technique is employed on multiple physically consecutive sectors on the same track to increase the compression efficiency. It results to reduction in read latency overhead by enabling a stronger ECC. This is termed as virtual sector compression. Because of overlapping writing in shingled recording, update-in-place feature fails to work. This makes the virtual sector compression practically possible. Two scenarios (fixed size and context aware virtual sector) are used to investigate the results of virtual sector compression. Intra-sector lossless compression results to the decrease in disk read latency.

Phasor measurement units (PMU) are used to monitor and regulate the power grid where the devices records globally synchronized measurements of voltage and current signals. It provides the respective phasor magnitude and angles at typical rate of 30 frames/s. To compress the phasor angle data, a new pre-processor is proposed. An adaptive high performance entropy encoder called golomb rice coding. This method involves two stages: In the first stage, preprocessing of data takes place. A new preprocessor based on frequency compensated differential encoding is proposed. In the second stage, golomb rice encoding is used. It is a lossless compression technique where the data transmission from space vessels takes place and higher throughput is achieved with limited memory and computational resources. Golomb encoding is designed to encode non-negative, integer valued signals in which probability value n decreases exponentially. It is very effective than existing coding techniques like Huffman coding and Arithmetic coding. Golomb rice coding is simpler to implement in hardware because it uses only few operators namely bitwise left, bitwise right, bitwise AND, OR, ++ and --. Next, mapping of values to GR codes does not require any lookup tables or trees/ GR compress 8 million phasor angles/second shows that it can be very useful for applications requires less delay. Data rate, compression rate and Error Propagation Rate (EPR) are used to evaluate the performance of this technique.

BWT COMPRESSION ALGORITHM ON SEISMIC DATA

The BWT [12] is a reversible block-sorting transform works on a series of data symbols to attain a permuted data sequence of similar symbols and an integer in {1, 2,..., n}. Let

$$BWT_n: X^n \rightarrow X^n \times \{1,2..,n\}$$

represents the n-dimensional BWT function and inverse of BWT is represented by

$$BWT_n^{(-1)}{}_n: X^n \times \{1,2..,n\} \rightarrow X^n$$

As the length of the series is identified from the source argument, the functionally transcript is removed and it gives

$$(y^n, u) = BWT(x_n)\ and\ BWT^{(-1)}((y^n, u) = x^n$$

where BWT_X represents the character and BWT_N represents the integer part of BWT. The forward BWT progresses by creating the all cyclic shifts of the uncompressed data string and arranging those cyclic shifts lexicographically. The BWT output has two parts. The first part is a length- string which gives the last character of each of the cyclic shifts. The next part is an integer which describes the position of the uncompressed data in the ordered list. For instance, BWT of the word “bananas” is given in [Fig. 1]. $BWT(banana) = bnnsaaaa, 4)$. While performing inverse BWT, reversible sequence transformation, the table should be reconstructed by the use of last column in the [Table 1], i.e. BWT output. Naturally, the reconstruction process of the table will be done in a column by column manner. Using the table construction, first column of the table is an ordered copy of the last column of the table. Hence, the *first* column can be reconstructed by replacing the alphabets on the list found in the last column. For reconstructing the second column, it is clearly shown that every row is a cyclic shift of every another row, and so last and first columns together gives a list of all successive pairs of symbols. Sorting this list of pairs produces the first and *second* column of the table. This process is repeated for third, fourth column and so on till all columns in the original table is reconstructed. The transform index represents the preferred row of the entire table. Inverse BWT is shown in [Fig. 2], $BWT^{(-1)}(bnnsaaa, 4) = banana.$

The steps involved in BWT are given below.

Step 1: Identify every cyclic sequence of the given data series
Step 2: Arrange the cyclic shift (rows) lexicographically
Step 3: The last column of the given array is given as output and row index of the input sequence.

The steps involved in inverse BWT are given below. For i=1, 2,.., n-1,

Step 1: Place column *n* in front of columns 1, 2,..., i-1
Step 2: Arrange the resulting length strings lexicographically
Step 3: Place the ordered list in the first I columns of the table.

	Step 1						
1	**b**	**a**	**n**	**a**	**n**	**a**	**s**
2	s	b	a	n	a	n	a
3	a	s	b	a	n	a	n
4	n	a	s	b	a	n	a
5	a	n	a	s	b	a	n
6	n	a	n	a	s	b	a
7	a	n	a	n	a	s	b

	Step 2							Step 3
1	a	n	a	n	a	s	b	b
2	a	n	a	s	b	a	n	n
3	a	s	b	a	n	a	n	n
4	**b**	**a**	**n**	**a**	**n**	**a**	**s**	s
5	n	a	n	a	s	b	a	a
6	n	a	s	b	a	n	a	a
7	s	b	a	n	a	n	a	a

Fig. 1: BWT of a sequence "banana".

	$(i=1)$					$(i=2)$				
	1	2	3			4	5	6		
1	b	a	a	⋯	b	ba	an	an	⋯	b
2	n	a	a	⋯	n	na	an	an	⋯	n
3	n	a	a	⋯	n	na	as	as	⋯	n
4	s	b	b	⋯	s	sb	ba	ba	⋯	s
5	a	n	n	⋯	a	an	na	na	⋯	a
6	a	n	n	⋯	a	an	na	na	⋯	a
7	a	s	s	⋯	a	as	sb	sb	⋯	a

	$(i=3)$			$(i=4)$		
	7	8	9	10	11	12
1	ban	ana	ana⋯b	bana	anan	anan⋯b
2	nan	ana	ana⋯n	nana	anas	anas⋯n
3	nas	asb	asb⋯n	nasb	asba	asba⋯n
4	sba	ban	ban⋯s	sban	bana	bana⋯s
5	ana	nan	nan⋯a	anan	nana	nana⋯a
6	ana	nas	nas⋯a	anas	nasb	nasb⋯a
7	asb	sba	sba⋯a	asba	sban	sban⋯a

	$(i=6, cont.)$
	18
1	ananasb
2	anasban
3	asbanan
4	**bananas**
5	nanasba
6	nasbana
7	sbanana

Fig. 2: IBWT of a sequence "(bnnsaaaa, 4)".

PERFORMANCE EVALUATION

To ensure the effectiveness of the BWT algorithm, its lossless compression performance is compared with 4 different, well-known compression algorithms namely Huffman coding, AC, LZW and LZMA.

Metrics

In the section, various metrics used to analyze the compression performance are discussed. The performance metrics are listed below: CR, CF, and BPC.

Compression Ratio (CR)
CR is defined as the ratio of the number of bits in the compressed data to the number of bits in the uncompressed data and is given in Eq. (1). A value of CR 0.62 indicates that the data consumes 62% of its original size after compression. The value of CR greater than 1 result to negative compression, i.e. compressed data size is higher than the original data size.

$$CR = \frac{\text{No. of bits in compressed data}}{\text{No. of bits in original data}} \quad (1)$$

Compression Factor (CF)
CF is a performance metric which is the inverse of compression ratio. A higher value of CF indicates effective compression and lower value of CF indicates expansion.

$$CF = \frac{\text{No. of bits in original data}}{\text{No. of bits in compressed data}} \quad (2)$$

bits per character (bpc)
bpc is used to calculate the total number of bits required, on average, to compress one character in the input data. It is defined as the ratio of the number of bits in the output sequence to the total number of characters in the input sequence.

$$bpc = \frac{\text{No. of bits in compressed data}}{\text{No. characters in the original data}} \quad (3)$$

Dataset description

For experimentation, a dataset which is used to identify seismic bumps larger than 104 J is chosen [13]. These data were collected from the mines in Poland with the time interval of 8 hours. The dataset consist of 2584 samples with two different classes 0 and 1. The value 0 indicates no earthquake hazard in the succeeding time interval and 1 indicates the occurrence of the earthquake hazard.

IIOAB JOURNAL
ISSN: 0976-3104

RESULTS AND DISCUSSION

To highlight the good characteristics of BWT based lossless compression algorithm for seismic dataset, it is compared with 4 states of art approaches namely Huffman coding, AC, LZW and LZMA. A direct comparison is made with the results of existing methods using the same dataset. Table 1 summarizes the experiment results of compression algorithms based on three compression metrics such as CR, CF and bpc. As evident from Table 1, the overall compression performance of BWT coding algorithm is significantly better than other algorithms on applied seismic bump dataset. It is observed that BWT coding achieves higher compression than existing methods. It is noted that Huffman coding and arithmetic coding produces poor results than BWT, LZW and LZMA.

The compression performance on applied dataset files reveals some interesting facts that the compression algorithms perform extremely different based on the nature of applied dataset. The existing LZW and LZMA method achieves better compression performance than Huffman coding and AC, but it fails to give performs well than BWT algorithm. Likewise, Huffman and Arithmetic coding also produce appropriately equal compression performance. This is due to the fact that the efficiency of the arithmetic coding is always better or at least identical to a Huffman code. Similar to Huffman coding, Arithmetic coding also tries to calculate the probability of occurrence of particular symbols and to optimize the length of the necessary code. It achieves an optimum code which exactly corresponds to the theoretical specifications of the information theory. A minor degradation results from inaccuracies, because of the correction operations for the interval division.

Table 1: Performance comparison of various compression methods on seismic data

Dataset	Compression techniques	Compression Ratio (CR)	Compression Factor (CF)	bits per character (bpc)
Seismic dataset	BWT	0.167	5.962	1.342
	LZMA	0.193	5.174	1.546
	Huffman	0.389	2.568	3.114
	LZW	0.285	3.512	2.277
	Arithmetic	0.383	2.608	3.066

On the other side, Huffman coding generates rounding errors because of its code length and is limited to multiples of a bit. The variation from the theoretical value is more than the inaccuracy of arithmetic coding. In overall, BWT coding results in effective compression than existing methods. Generally, dictionary based coding approaches find useful in situations where the original data to be compressed involves repeated patterns. LZMA extends LZW with the range encoding technique which enables to produce significantly higher compression than LZW. Interestingly, LZMA achieves significantly better compression with the compression ratio of 0.193 at the bit rate of 1.546 bpc. But, BWT achieves higher compression with the compression ratio of 0.167 at the bit rate of 1.342bpc.

CONCLUSION

This paper employs a Burrows Wheeler Transform (BWT) coding to compress seismic data. BWT is a block sorting compression which rearranges the character string into runs of identical characters. Extensive experiments were performed using real world seismic bump dataset. To ensure the effectiveness of lossless BWT coding, it is compared with 4 well-known compression algorithms namely Lempel Ziv Markov Algorithm (LZMA), Huffman coding, Lempel-Ziv-Welch (LZW) coding and Arithmetic coding (AC). By comparing the compression performance of BWT coding with existing methods, BWT achieves significantly better compression with a compression ratio of 0.167 at the bit rate of 1.342 bpc.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] M Bilen, AH Işık, T Yiğit. [2015] Seismic hazard prediction with classification of seismic pulses, International Burdur Earthquake & Environment Symposium (IBEES2015) Burdur, Turkey. 41-48.
[2] E Celik, M Atalay,H Bayer. [2014] Earthquake prediction using seismic bumps with Artificial Neural Networks and Support Vector Machines, Signal Processing and Communications Applications Conference, Trabzon, Turkey. 730-733.
[3] RW Hamming. [1986] Coding and information theory.
[4] Huffman DA [1952]. A Method for the Construction of Minimum-Redundancu Codes. 1098–1102.
[5] Vitter Jeffrey S. [1987] Design and Analysis of Dynamic Huffman Coding. 34: 825–845SW Drost, N Bourbakis. [2001] A Hybrid system for real-time lossless image compression. Microprocess. Microsyst. 25: 19–31. doi:10.1016/S0141-9331(00)00102-2
[6] SW Drost, N Bourbakis. [2001] A Hybrid system for real-time lossless image compression. Microprocess. Microsyst. 25: 19–31. doi:10.1016/S0141-9331(00)00102-2
[7] DA Huffman. [1952] A Method for the Construction of Minimum-Redundancu Codes. 1098–1102.
[8] H Witten Ian, M Neal Radford, G Cleary John. [1987] Arithmetic coding for DC. *Commun. ACM.* 30(6): 520–540.
[9] J Ziv. A Lempel. [1977] A Universal Algorithm for DC. *IEEE Trans. Inf. Theory*, 23(3): 337–343.
[10] J Ziv, A Lempel. [1978] Iz78.pdf. IEEE. 530–536.
[11] TA Welch. [1984] A technique for high-Performance DC. *IEEE*. 8–19.
[12] M Burrows, D Wheeler. [1994] A block-sorting lossless DC algorithm. *DC*. 124:18.
[13] https://archive.ics.uci.edu/ml/datasets/seismic-bumps

ISSN: 0976-3104
SPECIAL ISSUE: Computer Science

ARTICLE

A SIMPLE LOSSLESS COMPRESSION ALGORITHM IN WIRELESS SENSOR NETWORKS: AN APPLICATION OF WIND PLANT DATA

J. Uthayakumar[1*], T Vengattaraman[1], J. Amudhavel[2]

[1]*Department of Computer Science, Pondicherry University, Puducherry, INDIA*
[2]*Department of Computer Science and Engineering, KL University, Andhra Pradesh, INDIA*

ABSTRACT

Background: *Wireless Sensor Network (WSN) consists of numerous sensor nodes and is deeply embedded into the real world for environmental monitoring. In the last decades, a large amount of wind plant data is being generated as the interest in renewable sources is growing day by day. Data Compression (DC) techniques are commonly used to reduce the amount of data transmission. As the lossless compression produces reconstructed data same as original data, it is highly useful in applications where accuracy plays a major role.* ***Methods:*** *In this paper, a lossless compression algorithm is devised to compress the data generated in the wind plant monitoring and operation. Lempel Ziv Markov-chain Algorithm (LZMA) is employed to compress the wind plant data in WSN. LZMA algorithm compresses the wind plant data effectively. The sensor node in WSN runs LZMA and sends the compressed data to Base Station (BS).* ***Results:*** *Extensive experiments were performed using real world wind plant dataset such as Eastern dataset and Western dataset. To ensure the effectiveness of LZMA algorithm, it is compared with 3 well-known compression algorithms namely Burrows Wheeler Transform (BWT), Huffman coding and Arithmetic coding (AC).* ***Conclusions:*** *By comparing the compression performance of LZMA method with existing methods, LZMA achieves significantly better compression with a compression ratio of 0.107 (Eastern dataset) and 0.099 (Western dataset) at the bit rate of 0.855 bpc (Eastern dataset) and 0.798 bpc (Western dataset) respectively.*

KEY WORDS
Data Compression
LZMA algorithm
Wind plant data
Wireless Sensor Network

Received: 5 June 2017
Accepted: 25 July 2017
Published: 18 Sept 2017

***Corresponding Author**
Email:
uthayresearchscholar@gmail.com
Tel.: +91 9677583754

INTRODUCTION

The recent advancement in wireless networks and Micro-Electro-Mechanical-System (MEMS) leads to the development of low cost, compact and smart sensor nodes. WSN is randomly deployed in the sensing field to measure physical parameters such as temperature, humidity, pressure, vibration, etc. WSN is widely used in tracking and data gathering applications include surveillance (indoor and outdoor), healthcare, disaster management, habitat monitoring, etc. A sensor node is built up of four components namely transducer, microcontroller, battery, and transceiver. The sensor nodes are constrained in energy, bandwidth, memory and processing capabilities. As the sensor nodes are battery powered and are usually deployed in the harsh environment, it is not easy to recharge or replace batteries. The lifetime of WSN can be extended in two ways: increasing the battery storage capacity and effectively utilizing the available energy. The way of increasing the battery capacity is not possible in all situations. So, the effective utilization of available energy is considered as an important design issue. Several researchers observed that a large amount of energy is spent for data transmission when compared to sensing and processing operation. This study reveals that the reduction in the amount of data transmission is an effective way to achieve energy efficiency. Data transmission is the most energy consuming task due to the nature of strong temporal correlation in the sensed data. DC is considered as a useful approach to eliminate the redundancy in the sensed data. In WSN, the sensor node runs the compression algorithm and the compressed data will be forwarded to BS.

On the other side, wind plant data plays a major role in various data driven operations like integration studies and wind plant operation [1]. The wind plant data set holds many synchronized time sequences of wind power and wind speed in various locations. Generally, the wind plant dataset are larger in size, for example, 24 terabytes of data when archived in netCDF format [2]. The easier way to handle huge amount of data is to compress them. To reduce the amount of data being stored or transmitted, DC techniques has been proposed. It is applicable to compress text, images, audio or video. The data can be alphanumeric characters in a text document or numbers which represent the samples in an audio or image waveforms or series of numbers created by some processes, etc. DC can also be termed as an efficient way of representing the data in its compact manner. The compact form can be achieved by the recognition and utilization of patterns exists in the data. The evolution of DC starts with the Morse code, developed by Samuel Morse in the year 1838 [3]. Morse code is used in telegraph to compress letters. It uses smaller sequences to represent letters which occurs more frequently and thereby minimizing the message size and transmission time. This basic principle is employed in Huffman coding [4]. Nowadays, DC techniques are very essential in most of the real time applications like medical imaging, satellite imaging to Wireless Sensor Networks (WSN) etc. DC became important for utilization of available resources effectively. Without DC, it is very difficult and sometimes impossible to store or communicate huge amount of data files.

IIOAB JOURNAL
ISSN: 0976-3104

Basically, DC techniques falls under two categories: lossless compression and lossy compression. As the name implies, lossless compression refers that no loss of information i.e. the reconstructed data is exactly same as original data. It is used in situations where the loss of information is unacceptable. Example: Text, medical imaging, satellite imaging, etc. Lossless compression is generally used to compress text. As loss of information is intolerable in text compression, compressing text using lossless compression is mandatory. It achieves the compression ratio of 2:1 to 8:1 [5]. In some situations, the reconstructed data need not to be perfectly matched with the original data and the approximation of original data is also acceptable. In those situations, lossy compression is used which involves with some loss of information within the acceptable level. This leads to higher compression ratio when compared to lossless compression. Example: image, audio and video. Lossy compression is commonly used to compress images as the loss of image quality is tolerable. The main principle behind image compression is the correlation of pixels. The neighboring pixels of an image are high correlated. These redundant details are eliminated by computing a less correlated representation of the image. Lossy compression achieves the compression ratio of 100:1 to 200:1 based on the data to be compressed. It achieves higher compression ratio at a cost of loss in quality.

Only few researches have been made on the compression of wind plant dataset. Usually, they are compressed within the process information application with the help of dead-band and swinging door [6]. As they are lossy compression techniques, the same data cannot be obtained after the compression process. It leads to the development of lossless compression of wind plant data in the real world. Once the wind plant data has been compressed, it can be useful in several ways. In WSN, the sensor node executes the compression algorithm to compress the wind plant data. Then, the compressed data will be transmitted to BS. It can be easily shared, consumes less storage with less management cost. The absence of effective lossless compression algorithm for wind plant data motivated us to perform this work.

Contribution of the paper

The contribution of the paper is summarized as follows: (i) A lossless LZMA compression algorithm is used to compress wind plant data in WSN (ii) Two wind plant dataset from NREL (Eastern and Western Dataset) is used, and (iii) LZMA results are compared with 3 well-known compression algorithms namely Huffman coding, AC, and BWT interms of Compression Ratio (CR), Compression Factor (CF) and bits per character (bpc).

Organization of the paper

The rest of the paper is organized as follows: Section 2 explains the different types of classical DC techniques in detail. Section 3 presents the LZMA compression algorithm for wind plant dataset in WSN. Section 4 explains the performance evaluation in both dataset. Section 5 concludes with the highlighted contributions, future work, and recommendations.

RELATED WORK

DC is an efficient way to reduce the amount of data being stored or transmitted. DC compression techniques have been presented in [16]. The popular coding methods are Huffman coding, Arithmetic coding, Lempel Ziv coding, Burrows-Wheeler transform (BWT), RLE, Scalar and vector quantization.

Huffman coding [7] is the most popular coding technique which effectively compresses data in almost all file formats. It is a type of optimal prefix code which is widely employed in lossless DC. It is based on two observations: (1) In an optimum code, the frequent occurrence of symbols is mapped with shorter code words when compared to symbols that appear less frequently. (2) In an optimum code, the least frequent occurrence of two symbols will have the same length. The basic idea is to allow variable length codes to input characters depending upon the frequency of occurrence. The output is the variable length code table for coding a source symbol. It is uniquely decodable and it consists of two components: Constructing Huffman tree from input sequence and traversing the tree to assign codes to characters. Huffman coding is still popular because of its simpler implementation, faster compression and lack of patent coverage. It is commonly used in text compression.

AC [8] is an another important coding technique to generate variable length codes. It is superior to Huffman coding in various aspects. It is highly useful in situations where the source contains small alphabets with skewed probabilities. When a string is encoded using arithmetic coding, frequent occurring symbols are coded with lesser bits than rarely occurring symbols. It converts the input data into a floating point number in the range of 0 and 1. The algorithm is implemented by separating 0 to 1 into segments and the length of each segment is based on the probability of each symbol. Then the output data is identified in the respective segments based on the symbol. It is not easier to implement when compared to other methods. There are two

IIOAB JOURNAL
ISSN: 0976-3104

versions of arithmetic coding namely Adaptive Arithmetic Coding and Binary Arithmetic Coding. A benefit of arithmetic coding than Huffman coding is the capability to segregate the modeling and coding features of the compression approach. It is used in image, audio and video compression.

Dictionary based coding approaches find useful in situations where the original data to be compressed involves repeated patterns. It maintains a dictionary of frequently occurring patterns. When the pattern comes in the input sequence, they are coded with an index to the dictionary. When the pattern is not available in the dictionary, it is coded with any less efficient approaches. The Lempel-Ziv algorithm (LZ) is a dictionary-based coding algorithm commonly used in lossless file compression. This is widely used because of its adaptability to various file formats. It looks for frequently occurring patterns and replaces them by a single symbol. It maintains a dictionary of these patterns and the length of the dictionary is set to a particular value. This method is much effective for larger files and less effective for smaller files. For smaller files, the length of the dictionary will be larger than the original file. The two main versions of LZ were developed by Ziv and Lempel in two individual papers in 1977 and 1978, and they are named as LZ77 [9] and LZ78 [10]. These algorithms vary significantly in means of searching and finding matches. The LZ77 algorithm basically uses sliding window concept and searches for matches in a window within a predetermined distance back from the present position. Gzip, ZIP, and V.42bits use LZ77. The LZ78 algorithm follows a more conservative approach of appending strings to the dictionary.

LZW is an enhanced version of LZ77 and LZ78 which is developed by Terry Welch in 1984 [11]. The encoder constructs an adaptive dictionary to characterize the variable-length strings with no prior knowledge of the input. The decoder also constructs the similar dictionary as encoder based on the received code dynamically. The frequent occurrence of some symbols will be high in text data. The encoder saves these symbols and maps it to one code. Typically, an LZW code is 12-bits length (4096 codes). The starting 256 (0-255) entries represent ASCII codes, to indicate individual character. The remaining 3840 (256-4095) codes are defined by an encoder to indicate variable-length strings. UNIX compress, GIF images, PNG images and others file formats use LZW coding.

BWT [12] is also known as block sorting compression which rearranges the character string into runs of identical characters. It uses two techniques to compress data: move-to-front transform and RLE. It compresses data easily to compress in situations where the string consists of runs of repeated characters. The most important feature of BWT is the reversibility which is fully reversible and it does not require any extra bits. BWT is a "free" method to improve the efficiency of text compression algorithms, with some additional computation. It is s used in bzip2. A simpler form of lossless DC coding technique is RLE. It represents the sequence of symbols as runs and others are termed as non-runs. The run consists of two parts: data value and count instead of original run. It is effective for data with high redundancy.

A simpler form of lossless DC coding technique is Run Length Encoding (RLE). It represents the sequence of symbols as runs and others are termed as non-runs. The run consists of two parts: data value and count instead of original run. It is effective for data with high redundancy. For example, “ABCABBBBBC” is the input sequence, then the starting 4 letters are assumed as non-run with length 4, and the next 4 letters are assumed as a run with length 5 while B is repeated 5 times. RLE identifies the run of input file, saves the symbol and length of each run. These runs are used to compress the input file and non-runs are kept uncompressed. It is not effective in cases where the redundancy is low and leads to increase in original file size. It is ineffective for less redundant data and leads to increase in compressed file size greater than original file size. Input: YYBCCECCDESSEEEEER and its Output: 2Y1B2C1E2C1D1E2S5E1R. Though RLE is simpler to implement and faster, it fails to achieve better compression when compared to other algorithms. It is used in line drawing, animation and graphic icons, fax, palette-based images like textures.

Quantization is an easiest way to represent larger set of values to a smaller set. The quantizer input can be either scalar or vector. When the input is scalar, then it is called scalar quantizer. When the input is vector, then it is called vector quantizer. Scalar quantization is the simpler and commonly used lossy compression. It is a process of mapping an input value x to a number of output values. Vector quantization (VQ) is a different type of quantization, which is typically implemented by choosing a set of representatives from the input space, and then mapping all other points in the space to the closest representative. Initially, the source output is grouped into blocks or vectors. These vectors are used as the input to quantizer. In both sides of encoder and decoder, a codebook (set of L-dimensional vectors) is available. The vectors in the codebook are termed as code-vectors and each code vector is assigned to an index value. The elements of this code vector are the quantized values of the source output. In order to report the decoder about which code vector was found to be the closest to the input vector, the index value of the code-vector is transmitted. As the decoder also has same codebook, the code vector can be retrieved from its binary index.

IIOAB JOURNAL
ISSN: 0976-3104

Lossless compression techniques are widely used in text compression as loss of information is not tolerable while compressing text. A novel data compression technique is proposed to compress general data using a logical truth table (Mahmud, 2012). Here, the two bits of the data are represented by only one bit. Using this method, the data bits can be represented by its half bits only. i.e. 128 bits can be represented by 64 bits; 64 bits can be represented by 32 bits; 1GB data can be represented by 512MB and so on. The logic truth table like two inputs and four combinations are given below.
IF A=0 and B=0 then Z=0
IF A=0 and B=1 then Z=Ô
IF A=1 and B=1 then Z=1
IF A=1 and B=0 then Z=î

First, input data is tested to identify whether it is odd or even. If it is odd, extra bit is added at the last (0 is added when the last bit is 0 and vice versa). Then the data is split into two bits and truth table is applied to compress the data. Traditionally, there are two levels to represent digital data but proposed method uses four levels. The proposed method compresses the 20 bit data into 10 bits. This method uses Compression Ratio and Compression Factor as performance metrics. The author fails to compare the proposed method with existing method. It can be employed in wired and wireless scenarios.

Although some of common compression techniques can be applicable to some preferred applications, the concentration has been on the technique rather than on the application. However, there are certain techniques for which it is impossible to separate the technique from the application. This is because that several techniques depend on the properties or characteristics of the application. Therefore, some compression techniques were developed with focus on particular applications.

A new database compression method is proposed in large databases using association rule mining. Association rule mining from a database can identify frequent item sets in the database relations. The extracted association rules represent the repeated item sets in the database and they are compressed to decrease the file size. There are five steps in database compression: i. An effective association rule mining method using a tree called Class Inheritance Tree (CIT) is employed, ii. Collection of compression rules from strong association rules, iii. Compressed space for each compression rule is calculated, iv. A heuristic method is employed to resolve the conflicts of compression rules and v. Compression ratio is computed.

To reduce the Inter Track Interference (ITI) read latency of Shingled Magnetic Recording, two lossless compression mechanisms are used. It is a known fact that most of the files stored in disks are lossless compressible. When a sector of user data can be losslessly compressed to a particular level, free storage space will become available to store ECC redundancy. This extra space can be used for storing a stronger ECC than normal ECC for the current sector. This leads to decrease the probability of reading one or multiple neighboring tracks for explicit ITI compensation. Lempel-Ziv-Stores-Symanski (LZSS) algorithm is used for compressing several file types stored in hard disks. This method uses intra-sector data compression method to minimize the read latency is shingled recording. Here, lossless data compression technique is employed on multiple physically consecutive sectors on the same track to increase the compression efficiency. It results to reduction in read latency overhead by enabling a stronger ECC. This is termed as virtual sector compression. Because of overlapping writing in shingled recording, update-in-place feature fails to work. This makes the virtual sector compression practically possible. Two scenarios (fixed size and context aware virtual sector) are used to investigate the results of virtual sector compression. Intra-sector lossless compression results to the decrease in disk read latency.

Phasor measurement units (PMU) are used to monitor and regulate the power grid where the devices records globally synchronized measurements of voltage and current signals. It provides the respective phasor magnitude and angles at typical rate of 30 frames/s. To compress the phasor angle data, a new pre-processor is proposed. An adaptive high performance entropy encoder called golomb rice coding. This method involves two stages: In the first stage, preprocessing of data takes place. A new preprocessor based on frequency compensated differential encoding is proposed. In the second stage, golomb rice encoding is used. It is a lossless compression technique where the data transmission from space vessels takes place and higher throughput is achieved with limited memory and computational resources. Golomb encoding is designed to encode non-negative, integer valued signals in which probability value n decreases exponentially. It is very effective than existing coding techniques like Huffman coding and Arithmetic coding. Golomb rice coding is simpler to implement in hardware because it uses only few operators namely bitwise left, bitwise right, bitwise AND, OR, ++ and --. Next, mapping of values to GR codes does not require any lookup tables or trees/ GR compress 8 million phasor angles/second shows that it can be very useful for applications requires less delay. Data rate, compression rate and Error Propagation Rate (EPR) are used to evaluate the performance of this technique.

IIOAB JOURNAL
ISSN: 0976-3104

LZMA COMPRESSION ALGORITHM ON WIND PLANT DATA

LZMA compression is used to compress real time data generated rapidly. The overall operation is shown in [Fig. 1]. The LZMA algorithm compresses the wind power and speed values to reduce the size of the data. LMZA is the modified version of Lempel-Ziv algorithm to achieve higher CR [13]. It is a lossless DC algorithm based on the principle of dictionary based encoding scheme. LZMA utilizes the complex data structure to encode one bit at a time. It uses a variable length dictionary (maximum size of 4 GB) and is mainly used to encode an unknown data stream. It is capable of compressing the data generated at a rate of 10-20 Mbps in a real-time environment. Though it uses larger size dictionary, it still achieves the same decompression speed like other compression algorithms.

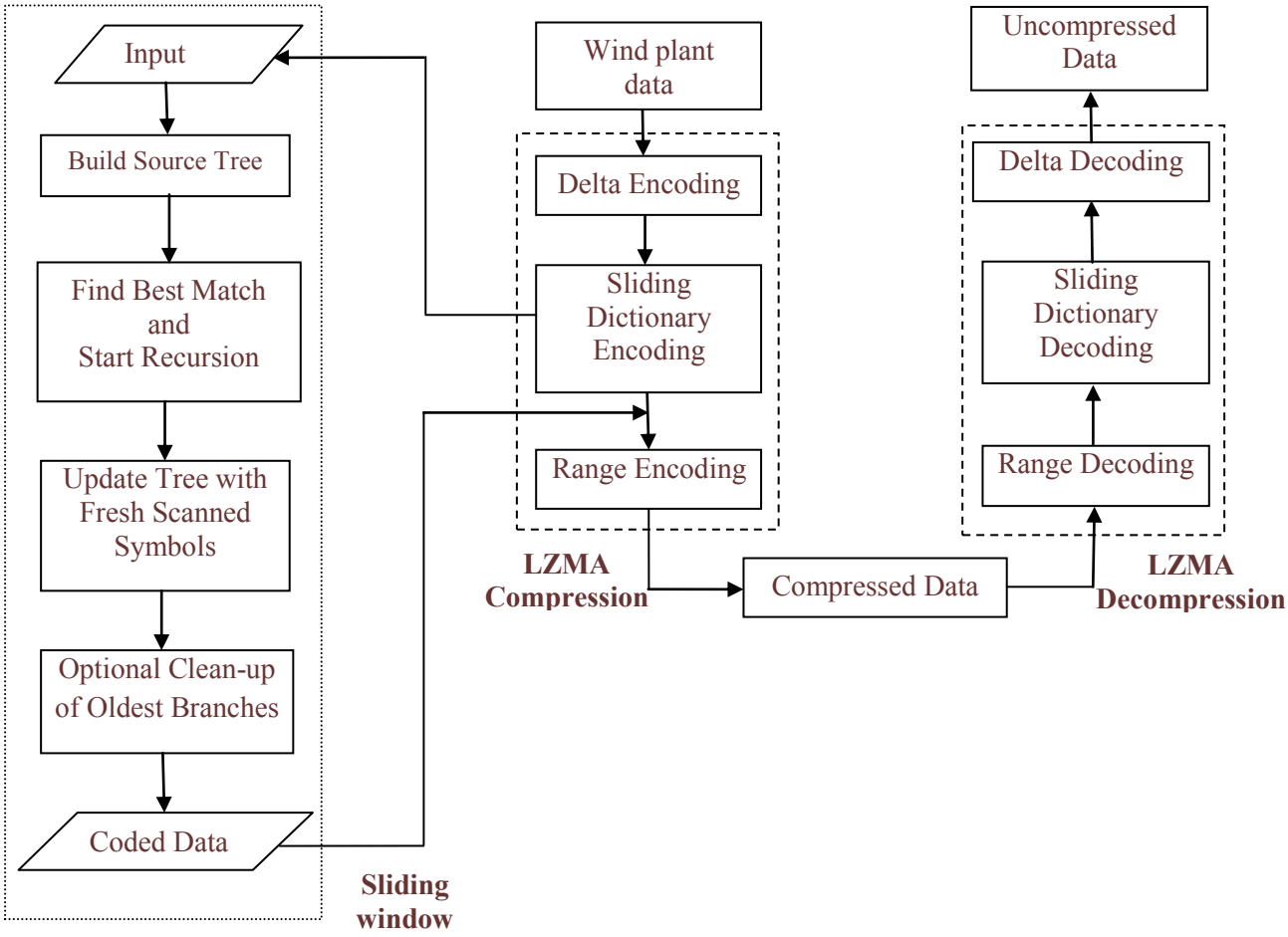


Fig. 1: Overall operation of LZMA on wind plant data.

LZ77 algorithm encodes the byte sequence from existing contents instead of the original data. When no identical byte sequence is available in the existing contents, the address and sequence length is fixed as '0' and the new symbol will be encoded. LZ77 also employs a dynamic dictionary to compress unknown data by the help of sliding window concept.

LZMA extends the LZ77 algorithm by adding a Delta Filter and Range Encoder. The Delta Filter alters the input data stream for effective compression by the sliding window. It stores or transmits data in the form of differences in sequential data instead of complete file. The output of the first-byte delta encoding is the data stream itself. The subsequent bytes are stored as the difference between the current and its previous byte. For continuously varying real time data, delta encoding makes the sliding dictionary more efficient. For example, consider a sample input sequence as 2,3,4,6,7,9,8,7,5,3,4 7. The input sequence is encoded with LZMA technique and the encoded output sequence is 2, 1, 1, 2, 1, 2,-1,-1,-2,-2, 1. So, the number of symbols in the input sequence is 8 and the number of symbols in output sequence is 4.

IIOAB JOURNAL
ISSN: 0976-3104

Static and adaptive dictionary are the commonly used dictionaries. The static dictionary uses the fixed entries and constants based on the application of the text. Adaptive dictionaries take the entries from the text and generate on run time. A search buffer is employed as a dictionary and the buffer size is chosen based on the implementation parameters. Patterns in the text are assumed to occur within range of the search buffer. The offset and length are individually encoded, and a bit-mask is also separately encoded. Usage of an appropriate data structure for the buffer decreases the search time for longest matches. Sliding Dictionary encoding is comparatively tedious than decoding as it requires to identify the longest match. Range encoder encodes all the symbols of the message into a single number to attain better CR. It efficiently deals with probabilities which are not the exact powers of two.

The steps involved in range encoder are listed below.

- Given a large-enough range of integers, and probability estimation for the symbols.
- Divide the primary range into sub-ranges where the sizes are proportional to the probability of the symbol they represent.
- Every symbol of the message is encoded by decreasing the present range down to just that sub-range which corresponds to the successive symbol to be encoded.
- The decoder should have same probability estimation as encoder used, which can either be sent in advance or derived from already transferred data.

PERFORMANCE EVALUATION

To ensure the effectiveness of the LZMA algorithm while compressing wind plant data, its lossless compression performance is compared with 3 different, well-known compression algorithms namely Huffman coding, AC, and BWT.

Metrics

In the section, various metrics used to analyze the compression performance are discussed. The performance metrics are listed below: CR, CF, and bpc.

Compression Ratio (CR)
CR is defined as the ratio of the number of bits in the compressed data to the number of bits in the uncompressed data and is given in Eq. (1). A value of CR 0.62 indicates that the data consumes 62% of its original size after compression. The value of CR greater than 1 result to negative compression, i.e. compressed data size is higher than the original data size.

$$CR = \frac{\text{No. of bits in compressed data}}{\text{No. of bits in original data}} \quad (1)$$

Compression Factor (CF)
CF is a performance metric which is the inverse of compression ratio. A higher value of CF indicates effective compression and lower value of CF indicates expansion.

$$CF = \frac{\text{No. of bits in original data}}{\text{No. of bits in compressed data}} \quad (2)$$

bits per character (bpc)
bpc is used to calculate the total number of bits required, on average, to compress one character in the input data. It is defined as the ratio of the number of bits in the output sequence to the total number of character in the input sequence.

$$BPC = \frac{\text{No. of bits in compressed data}}{\text{No. character in the original data}} \quad (3)$$

Dataset description

For experimentation, two publicly available wind pant dataset are used. The data is collected by National Renewable Energy Laboratory (NREL). The two major sources of data is used in this work are NREL Eastern [14] and NREL Western dataset [15]. The data were collected by AWS Truepower and 3-TIER. The Eastern dataset holds time sequences of 1326 hypothetical wind plant of different locations, capacities and wind turbine power curves, the capacity of the wind plant are in the range of 100 to 1435MW with a median

IIOAB JOURNAL
ISSN: 0976-3104

capacity of 370MW. These data were archived at an accuracy of 100kW and 0.001m/s. The Western dataset holds time series of 32000 hypothetical wind plant of 1.2 million different locations, capacities and wind turbine power curves, with the same capacity of 30 MW. These data were archived at an accuracy of 1kW and 0.01m/s.

RESULTS AND DISCUSSION

To highlight the good characteristics of LZMA based lossless compression algorithm for wind plant data, it is compared with 3 states of art approaches namely Huffman coding, AC and BWT coding. A direct comparison is made with the results of existing methods using the same set of 2 datasets. Table 1 summarizes the experiment results of compression algorithms based on three compression metrics such as CR, CF and bpc. As evident from [Table 1], the overall compression performance of LZMA algorithm is significantly better than other algorithms on both two dataset. It is observed that LZMA algorithm achieves slightly higher compression on Western dataset than Eastern dataset. It is also noted that Huffman coding and arithmetic coding produces poor results than BWT. The compression performance on 2 different dataset files reveals some interesting facts that the compression algorithms perform extremely different based on the nature of applied dataset. The existing BWT method achieves better compression performance than Huffman coding and AC, but it fails to give performs well than LZMA algorithm. Likewise, Huffman and Arithmetic coding also produce appropriately equal compression performance. This is due to the fact that the efficiency of the arithmetic coding is always better or at least identical to a Huffman code. Similar to Huffman coding, Arithmetic coding also tries to calculate the probability of occurrence of particular certain symbols and to optimize the length of the necessary code. It achieves an optimum code which exactly corresponds to the theoretical specifications of the information theory. A minor degradation result from inaccuracies, because of correction operations for the interval division.

Table 1: Performance comparison of various compression methods on wind plant data

Dataset	Compression techniques	Compression Ratio (CR)	Compression Factor (CF)	bits per character (bpc)
Eastern Dataset	LZMA	0.107	9.356	0.855
	Huffman	0.566	1.766	4.528
	BWT	0.133	7.545	1.060
	Arithmetic	0.569	1.789	4.470
Western Dataset	LZMA	0.099	10.026	0.798
	Huffman	0.665	1.769	4.520
	BWT	0.109	9.151	0.874
	Arithmetic	0.559	1.787	4.476

On the other side, Huffman coding generates rounding errors because of its code length and is limited to a multiples of bit. The variation from the theoretical value is more than the inaccuracy of arithmetic coding. In overall, LZMA results in effective compression than existing methods. Generally, dictionary based coding approaches find useful in situations where the original data to be compressed involves repeated patterns. LZMA extends LZW with the range encoding technique which enables to produce significantly higher compression than LZW. Interestingly, LZMA achieves significantly better compression with the compression ratio of 0.107 (Eastern dataset) and 0.099 (Western dataset) at the bit rate of 0.855 (Eastern dataset) and 0.798 (Western dataset) respectively.

CONCLUSION

This paper employs a Lempel Ziv Markov chain Algorithm (LZMA) lossless compression technique to compress the wind plant data. LZMA is a lossless DC algorithm which is well suited for real time applications. LZMA algorithm compresses the wind plant data effectively. Extensive experiments were performed using the real world wind plant dataset such as Eastern dataset and Western dataset. To ensure the effectiveness of LZMA

IIOAB JOURNAL
ISSN: 0976-3104

algorithm, it is compared with 3 well-known compression algorithms namely Burrows Wheeler Transform (BWT), Huffman coding and Arithmetic coding (AC). By comparing the compression performance of LZMA method with existing methods, LZMA achieves significantly better compression with a compression ratio of 0.107 (Eastern dataset) and 0.099 (Western dataset) at the bit rate of 0.855 bpc (Eastern dataset) and 0.798 bpc (Western dataset) respectively.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Ummels HHB, Tande JO, Estanqueiro A, et al. [2009] Design and operation of power systems with large amounts of wind power Final Report of IEA Task 25, VTT Research Notes.

[2] Potter CW, Lew D, McCaa J, Cheng S, Eichelberger S, Grimit E. [2008] Creating the dataset for the western wind and solar integration study (U.S.A)," in *Proc. 7th Int. Workshop on Large Scale Integration of Wind Power and on Transmission Networks for Offshore Wind Farms*, Madrid, Spain.

[3] Hamming RW. [1986] Coding and information theory.

[4] Vitter Jeffrey S. [1987]. Design and Analysis of Dynamic Huffman Coding. 34: 825–845.

[5] Drost SW, Bourbakis N. [2001] A Hybrid system for real-time lossless image compression. Microprocess. Microsyst. 25: 19–31. doi:10.1016/S0141-9331(00)00102-2

[6] Srisooksai T, Keamarungsi K, Lamsrichan P, Araki K. [2012] Practical DC in wireless sensor networks: A survey. *J Netw Comput Appl* 35 (1): 37–59.

[7] Huffman DA. [1952] A Method for the Construction of Minimum-Redundancu Codes. 1098–1102.

[8] Witten Ian H, Neal Radford M, Cleary John G. [1987] Arithmetic coding for DC. *Commun. ACM.* 30(6): 520–540.

[9] J Ziv. A Lempel. [1977] A Universal Algorithm for DC. *IEEE Trans. Inf. Theory*, 23(3): 337–343.

[10] Ziv J, Lempel A. [1978] lz78.pdf. IEEE. 530–536.

[11] Welch TA. [1984] A technique for high-Performance DC. *IEEE*. 8–19.

[12] Burrows M, Wheeler D. [1994] A block-sorting lossless DC algorithm. *DC*. 124:18.

[13] Tu Z, Zhang S. [2006] A Novel Implementation of JPEG 2000 Lossless Coding Based on LZMA. in *Proceedings of the Sixth IEEE International Conference Computer and Information Technology*.

[14] NREL, Wind Integration Datasets Aug. [2010] [Online]. Available: http://www.nrel.gov/wind/integrationdatasets/eastern/data.html

[15] NREL, Wind Integration Datasets Aug. [2009] [Online]. Available: http://www.nrel.gov/wind/integrationdatasets/western/data.html

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

LIST ALGORITHM: LINEAMENT IDENTIFICATION AND STOCKWORKS TARGETING WITHIN ASTER SATELLITE IMAGE

Sukumar M[1*], C Nelson Kennedy Babu[2]

[1]*Department of Information Technology, St.Peter's Institute of Higher Education & Research, St. Peter's University, Avadi, Chennai, Tamil Nadu, INDIA*

[2]*Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering, Navakkarai, Coimbatore, Tamil Nadu, INDIA*

ABSTRACT

Background: *Now-a-days, remote sensing has been widely used in the field of geology. The application of remotely sensed data for lineament interpretation in mineralogical environment is demonstrated here. Lineament investigation is classified based on the geological properties, and lineament parameters like its position, direction. Satellite images (or) Aerial images are used now-a-days to discriminate the lineaments based on the spatial variations in orientation and density. The main purpose of this work is to identify lineaments and target stockworks from the ASTER satellite image which contributes to understand about the geological faults over the study area.* ***Methods:*** *The major work for the extraction of lineaments and its intersections is implemented using the following steps: A) selection of the suitable band from the ASTER satellite image for lineament extraction and the following geospatial analysis. B) applying image preprocessing methods to denoise and enhance the contrast, edge information of lineaments C) Identifying the lineaments using the proposed Lineament Identification and Stockworks Targeting (LIST) algorithm D) Analyzing the following geospatial aspects: lineament length, density and stockworks.* ***Results:*** *The proposed LIST algorithm solved many of the existing problems encountered while analyzing lineaments and the result of LIST is compared with Mask based clustering and Fuzzy C Means clustering techniques which is mostly used by other researchers. LIST algorithm concentrates more on micro lineaments and stockworks area which has the higher potential for minerals existence. A total of 699 lineaments with total length of 17,561kms are extracted using LIST which is comparatively higher than the other two methods discussed here.* ***Conclusion:*** *the geospatial analysis result shows that the identification of geological lineaments over wide area is possible and detailed information for the purpose of exploring minerals and its prospective zones can be obtained using the proposed LIST algorithm. The knowledge of lineaments can be applied to the terrain surfaces to map/monitor the road networks, drainage density networks, and alignment of vegetation.*

KEY WORDS
Lineament, Stockworks, LIST, ASTER Satellite, Afar depression

Received: 15 May 2017
Accepted: 31 July 2017
Published: 20 Sept 2017

***Corresponding Author**
Email:
msukumar.btech@gmail.com
Tel.: +91-94867 37466

INTRODUCTION

The alignment of faults and fractures were analysed during the 19th century in Great Britain. Hobbs initiated the work [1] on "lineaments of the Atlantic border regions". He defined lineaments as "significant lines in the Earth's surface" and stated that "an essential lineaments as the top of ridges or boundaries of elevated areas, drainage lines, coastlines and limits of geological formation of petrographic rocks or vegetation line of crops" and then the definition is rephrased as that "significant lines in the landscape which reveal the hidden architecture". Most of the authors classified the lineaments based on its width, length and contrast.

Identifying lineaments from the satellite image [2, 3] can be done using two steps. At first, selecting suitable band for processing, applying edge enhancement methods, then extracting lineaments using GIS software. Various automatic tools and software's are used since only one software tool is not sufficient to process all the steps of analysis. Most of the studies utilised the following software tools. LINDEN [3], PCI Geomatica [4-6] software extracts lineaments using edge detection, thresholding and curve extraction steps. But this kind of extraction may detect non geologic lineaments (e.g.) manmade linear structures like road, pipelines and drainage networks etc., LINE module, ArcGIS, ArcMap, GEOrient [7, 8]. "PhotoLin" tool [9] is developed exclusively for IBM compatible PCs to extract lineaments from aerial, satellite and topographic images. For that, the given image is binarized, segmented using threshold to locate the features of interest and then the thinning algorithm is applied to generate the lineament map. Tonal / Contrast based methods are applied to differentiate lineaments. In Segment Tracing Algorithm lineaments are extracted based on the local variance of grey level and connect two line elements of same orientation. The main advantage of this method is that the capability to track continuous valleys and its shadows which look like two parallel lines in the lineament map of the study area.

An adaptive anisotropic kernel method [10] estimates the lineament density and distribution by adjusting the smoothing kernel based on the local spatial distribution, and then the global thresholding is applied to segment the density image. Whenever the number of objects is large, the natural way of counting the object is by using pointing (or) dotting the objects. Likewise, the spatial arrangement of the dots provides wealthy additional information on density analysis. Counting by detection method localises individual object instances in the image. Locating and counting the overlapping objects becomes trivial and hard. Instead of solving the hard problems of overlapping, Counting by regression method considers the global characteristics (histograms of image features) to detect the number of objects and this method cannot consider other information like the location and the population of the objects. Counting by segmentation methods [11, 12] combines both the methods which segment the image objects into separate clusters and then regresses from the global properties of each group to the overall number of objects in it. Various classification algorithms and its combinations are used to categorize the land use areas from the satellite image [13, 14].

IIOAB JOURNAL
ISSN: 0976-3104

Fracture Network Evaluation Program (FraNEP)[15] is developed to analyse various parameters of lineament such as lineament density, intensity, and length using scanline sampling and window sampling. FracSim 3D software package analyses the fracture lengths with the help of image histograms. The LINDENS tool estimates the lineament density. Some of the lineament analysis works [16, 17] focused on tectonic and geochemical estimation which determines the orientation of fault and fractures**.** Canny algorithm [18] with different window size extracts lineaments and its alteration zones. Evaluation of density and frequency of lineaments helps to update the fault / geological map of the study area. From the detailed review of these works results the following limitations and challenges exists. They are:

- Lineament exhibits different brightness and contrasts on its sides because of poor illumination.
- Global scale segmentation cannot be able to detect the micro lineaments.
- Discontinuity in the linear pattern even in the same fault occurs if edge detection is poor.
- Extraction of stockworks is hard; if the lineament object is over / under segmented.
- Difficult to update geologic lineament maps by field observations. Lack of up-to-date lineament maps.

The main objective of the current research work is to provide a Lineament and Stockworks extraction system. For that, a new algorithm named LIST (Lineament Identification and Stockworks Targeting) is proposed here to solve the above-listed problems which are encountered while analyzing lineaments. Finally visual interpretation of manually extracted lineaments [19] with the geologic map of Ethiopia and also the automatically extracted lineaments can be done to update the geologic lineament maps of an area. Macro, moderate and micro lineaments and its geospatial results along with the field evidence reveal firm evidence of geologic/tectonic scenario.

MATERIALS AND METHODS

ASTER image dataset

Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery of 30-m resolution [20] gives better results while analysing lineaments when compared to IRS 1D imagery because of improved spatial resolution. To classify the geologic lineaments from the High-resolution Multispectral image, ASTER satellite image is used. ASTER is a multispectral imager which covers a wide spectral region of the electromagnetic spectrum from the Visible Near Infra-Red (VNIR) to the Thermal Infra-Red (TIR). Recent years, the use of ASTER data in mineral prospecting application is increased because of it is freely available on the internet, the image covers a broad area (approx. 60 sq.kms), and different integral bands are highly sensitive to alteration minerals (i.e.) minerals known to surround target minerals. Lithologic and mineralogical units on the surface can be accurately mapped with this satellite image. 15m resolution of VNIR data is one of the best resolutions among all the multispectral satellite data available commercially.

Description of study area

The Afar is one of the northeastern states of Ethiopia [21]. The Afar Depression is an area of lowland plains split by fault blocks and dotted with shield volcanoes. It is the lowest valley region in Ethiopia and one of the lowest in Africa. Afar depression is surrounded by southern Red Sea rift on its northern side. In the eastern part, the Gulf of Aden Rift is spread through the Gulf of Tajura. The Ethiopian Plateau is located in the west. The Somalian Plateau covers the southern region. The Danakil Depression extends to its Northeast, and the south-east is Ali-Sabieh block. Southwest extension continues through the Main Ethiopian Rift to the East African Rift System. It is widely a desert scrubland with shallow salty lakes and long chains of volcanoes.

Geology of the study area

The geology and tectonics of the Afar Depression [22] is a plate tectonic triple junction where the spreading ridges that are forming the Red Sea and the Gulf of Aden emerge on land and meet the East African Rift. The Afar Depression is one of two places on Earth where a mid-ocean ridge can be studied on land, the other being Iceland. The Afar is slowly being pulled apart at a rate of 1-2 cm per year. This geologic feature is one of earth's great active volcanic areas [23]. Due to this volcanic activity the floor of the depression is composed of lava, mostly basalt. The continuous process of volcanism results in the occurrence of major minerals including potash, sulphur, salt, bentonite, and gypsum. Most of the region's mineral potential is found in Afar state. The overall structural trend across the Afar Depression is for the older larger border faults to be abandoned and for active faulting to be concentrated along narrow zones within the centre of wider rift systems. This is due to the decrease in the strength of the lithosphere with increasing strain rates. ASTER image of Afar region for this work was acquired on March 7, 2002, is located near 12.2 degrees north latitude and 41.9 degrees east longitude, and covers an area of 57 x 61.7 km (approximately 3500 sq.kms).

Geospatial analysis

Several researchers have conducted lineament length and density studies. Lineament density analysis can be done to identify or locate the weathered zone and also helps to find out the ground water potential area, landslides, and so forth. Lengthy lineaments are the zones having higher permeability and porosity [24]**.** Lineament density map is generated to describe the distribution of lineaments in a two-dimensional

plot. The image is split into equal sized polygons. The number of lineaments in each polygon is taken into account which shows the concentration of lineaments over the study area. The result of this analysis helps to know the densely populated lineaments area in the image. A new type of fractal plot is developed based on the fractal nature of lineaments. Optimized lineament density map is generated with the help of Optimal Cell Dimension (OCD) which depicts the point of intersection of two lines on the plot. A statistical analysis is being carried out to count lineaments and its intersection regions within the cell dimension. This results various density maps.

Methodology used

In this work, lineament extraction process comprises the following four steps: (i) image preprocessing (ii) lineament identification (iii) stockworks extraction and (iv) lineament / stockworks density map generation.

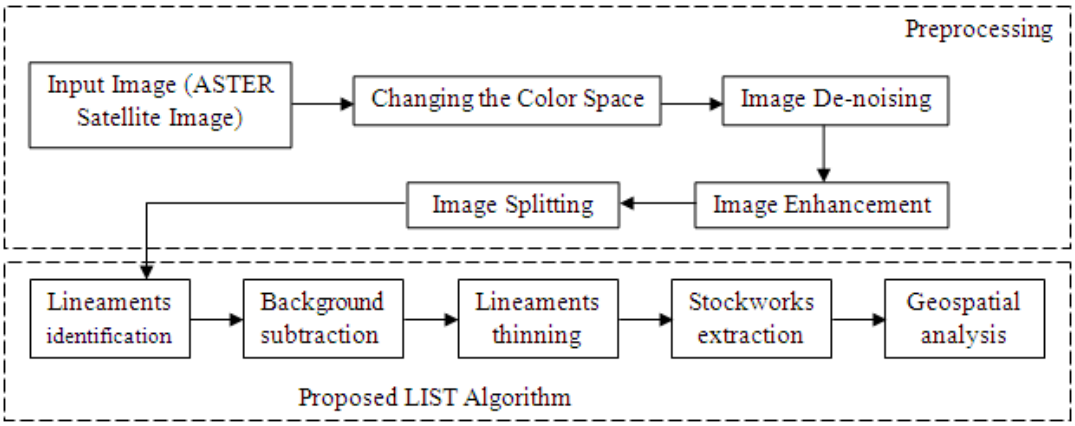


Fig. 1: Workflow diagram

Changing the colour space

Whenever an observer is moved further away from a scene, an illusion produced by the visual system is a mixing of colours. Initially, the viewer notices the difference between those colours; because these colours activate two different sensors in the retina. However, if the object is away from the scene, then these two colours activates the same sensor present in the retina, this enables the observer to see only one colour which composed of varying real colours in the scene. CIE L*a*b* (CIELAB) is a colour space specified by the International Commission on Illumination. It is the device independent colour space which denotes all the colours visible to the human eye. In this colour system, "L" shows the colour lightness, "A" describes the colours in between red/magenta and green, and "B" shows the colours in between yellow and blue

Denoising

The isotropic undecimated wavelet transform is suitable for planetary imaging. It decomposes the image into different scales. IUWT introduces a multi-resolution algorithm for detecting bright spots. To keep the important response of the filter to the desired feature, the denoising technique [25] uses hard thresholding value. Finally, the newly selected coefficient allows us to combine multiscale information to detect the spots. However, its performance is slightly poor in case of low-quality images, at that time, soft thresholding is used; instead of hard thresholding.

Contrast enhancement

One of the effective ways to enhance the image is by equalising the histogram values of the image. Histogram equalisation methods enhance the image fully i.e. it does not consider the contrast and brightness (intensity) values present in the image. It creates an undesirable effect while post processing the image. Brightness preserving dynamic fuzzy histogram equalisation [BPDFHE] technique equalises the image histogram by distributing the grey values present in the valley portions of the histogram. It clearly shows that no remapping of the histogram peaks takes place. This method is used in both grayscale and colour images.

- Compute and partition the histogram based on the “local maxima” value. To reduce the approximation errors, second order derivative is calculated from the fuzzy histogram.
- Global Histogram Equalization method is used to equalise every partitioned histogram dynamically based on the highest and lowest intensity values contained in the partitioned histogram.
- Normalising the brightness of the image.

IIOAB JOURNAL
ISSN: 0976-3104

Image splitting

Splitting the image is termed as the process of dividing the image into equal sized blocks. The basic structure is non-overlapping square blocks of size m x m. To identify the micro lineaments, this condition is applied to split the image, and after processing, the blocks of equal size are merged to form a whole image. If the block differs in size, the merging process terminates, and no further merge is possible.

- Split the input image into equal sized square blocks/grids.
- Before extracting the lineaments, the line support regions (i.e. clusters having the same intensity which depicts lineaments) are formed. Find the magnitude of the image and the direction of all the lineaments to avoid the boundary effect of each grid while joining the grids after processing.
- Local search of coherent components is characterized using two parameters: intensity and shape. The intensity similarity between the candidate pixel and its neighbourhood values are analysed. Thus the starting and ending point coordinates of the lineaments are detected

LIST (Lineament Identification and Stockworks Targeting)
Lineament Identification

- Eigenvectors can be used to formulate the lineaments of various angle projections from the grid image. In this step, scatter matrix is formed to check whether the pixel contributes to form a lineament or not and this represents the spatial relationship between the neighbouring pixels in an image.

$$sm = \begin{pmatrix} sm_{11} & sm_{12} \\ sm_{21} & sm_{22} \end{pmatrix}$$

- Consider 'm' is the number of pixels in the line, (A_i, B_i) is the coordinates of the i^{th} pixel on the line. Eight neighbourhood elements and 24 neighbourhood elements of an image are analysed to detect the lines efficiently.

$$A_m = \begin{bmatrix} -1,-1 & -1,0 & -1,1 \\ 0,-1 & & 0,1 \\ 1,-1 & 1,0 & 1,1 \end{bmatrix} \qquad B_m = \begin{bmatrix} -2,-2 & -2,-1 & -2,0 & -2,1 & -2,2 \\ -1,-2 & -1,-1 & -1,0 & -1,1 & -1,2 \\ 0,-2 & 0,-1 & & 0,1 & 0,2 \\ 1,-2 & 1,-1 & 1,0 & 1,1 & 1,2 \\ 2,-2 & 2,-1 & 2,0 & 2,1 & 2,2 \end{bmatrix}$$

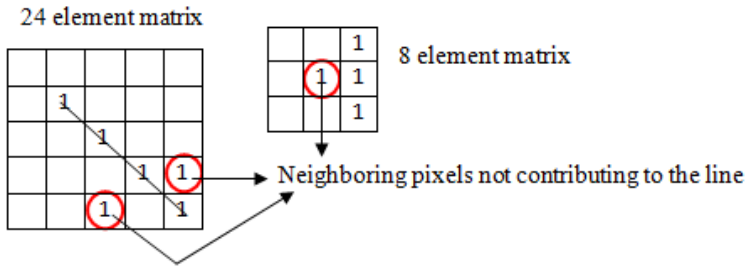


Fig. 2: Eight element matrix and twenty four element matrix.

- Now, the coordinate of the i^{th} pixel on the line is termed as

$$sm_{11} = \frac{1}{2}\sum_{i=1..n}(A_i - A_m)^2 \qquad sm_{22} = \frac{1}{2}\sum_{i=1..n}(B_i - B_m)^2$$

- Resultant large eigenvalue k_1 and small eigenvalue k_2 of scatter matrix is formulated as

$$k_1 = \frac{1}{2}\left[sm_{11} + sm_{22} + \sqrt{(sm_{11} - sm_{22})^2 + 4sm_{11}^2}\right]$$

$$k_2 = \frac{1}{2}\left[sm_{11} + sm_{22} - \sqrt{(sm_{11} - sm_{22})^2 + 4sm_{22}^2}\right]$$

- Calculate the lineament length and the number of lineaments.
- Classify the lineaments based on its length. In this work, lineaments are categorised into
  - Micro lineaments – lineament length less than 1 km
  - Moderate lineaments - lineament length greater than 1 km and less than 2 km.s
  - Macro lineaments - lineament length greater than 2 km.s

```
function Length_Analysis (input image, the length of the lineament)
if lineament length<threshold level
        Set colour code for micro lineament and increment its count
else if lineament length > threshold level
        Set colour code for macro lineament and increment its count
else
        Set color code for moderate lineament and increment its count
end
```

- Lineaments may fall/extend on more than one block. All the blocks are merged to extract the total lineaments of the geological area. This can be done by repeating the steps 3 to 7

Background subtraction & lineament thinning

- Background subtraction is a common approach is used to detect the object of interest in the foreground thus the background information is not needed for processing. After image preprocessing, object localisation is required for further processing. In this work, lineaments detected from the original image are highlighted alone by suppressing the background of the image.
- Subtract the background image from the lineaments detected image and increase the intensity of the lineaments image for further processing

$$B(x,y) = \frac{1}{N}\sum |L_i(x,y) - L_j(x,y)|$$

- Where N is the number of pixels in the image used as a scaling factor. L_i and L_j are the images before and after lineament extraction
- The relationship between the lineament pixel and its neighbourhood varies the lineament width while increasing the intensity after background subtraction which leads to the false stockworks mapping. This can be avoided by applying the morphological thinning operation on the extracted lineaments image.

$$A \oplus B = \max_{[j,k]\in B}\{a[m-j,n-k]+b[j,k]\}$$

where, A – Input image & B – Structural element

Stockworks extraction

- Set the cross point pixel as the cumulative successive difference between a pair of adjacent pixel belongs to 8 neighbourhood of point p.

$$c_n(p) = \frac{1}{2}\sum_{i=1}^{8} |val(p_i \bmod 8) - val(p_{i-1})|$$

- Where p_0, p_1 and p_2 are the pixels belonging to 8 neighbourhood of p. If the centre candidate pixel has only one one valued neighbour, then the pixel is a cross point termination pixel, and if the centre candidate pixel has three one valued neighbour, then the pixel is said to be a cross point pixel. If the centre candidate pixel has two one valued neighbour, it is the usual pixel not termed as the cross point.

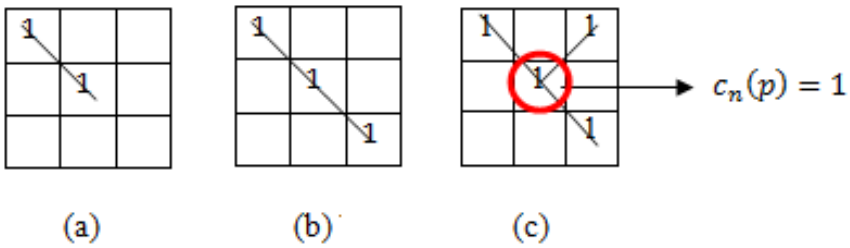


Fig. 3: (a) Lineament termination (b) usual line and (c) stockworks.

To map the density feature of the stockworks and the lineaments, the spectral similarity measure is calculated by measuring the angle between the spectral value of two samples k_i and k_j. This is insensitive to illumination effect because the angle between the two vectors is constant according to the length of the vector.

$$\Delta k = \arccos\left(\frac{k_i . k_j}{|k_i| . |k_j|}\right)$$

To calculate the angle deviation of the spectral values, the Gaussian function is defined as

$$F(\Delta k) = \exp\left[\frac{-1}{2}\left(\frac{\Delta k}{\sigma \Delta k}\right)^2\right]$$

$$|k| = \left(\frac{(|k_i| - |k_j|)}{\max(|k_i| - |k_j|)}\right) \quad \forall\, [0..1]$$

For similarity prediction,

$$k(w_i, w_j) = \max\{k_1[\Delta k], k_2[\Delta k]\} \quad \forall\, [0..1]$$

This highlights the spatially disjoint and spectrally similar areas in the given satellite image

IIOAB JOURNAL
ISSN: 0976-3104

RESULT & DISCUSSION

Geological lineaments are the faults, fractures and folds on the ground surface. The geospatial analysis of these lineaments and the stockworks were carried out in this work. This algorithm generates blocks that divides the study area and locate the lineaments from each block and also find the length and the density of the lineaments on those blocks. It also maps the stockworks and its density regions. [Fig.4(a)] shows the original input image of a part of afar region and 50 parts of the same region images are also taken into account for detecting lineaments and stockworks. Then the identified lineaments on those locations are superimposed on the input image after several processing steps and finally stockworks are extracted [Fig.4(b)] from the image. [Fig.5(a)] shows the geospatial analysis output of LIST algorithm. In that analysis, [Fig.5(b)], [Fig.5(c)] and [Fig.5(d)] shows various categories of lineaments based on its length – Macro lineaments, Moderate lineaments and Micro lineaments. [Fig.5(e)] depicts stockworks density map and [Fig.5(f)] depicts lineament density map, the density of the lineaments is categorized into three classes by different colors patterns. The denser the area of the lineaments seems to be the best prospective zone for mineralization. Likewise, various kinds of density regions of stockworks are generated to describe the varying quantity of minerals located on that area.

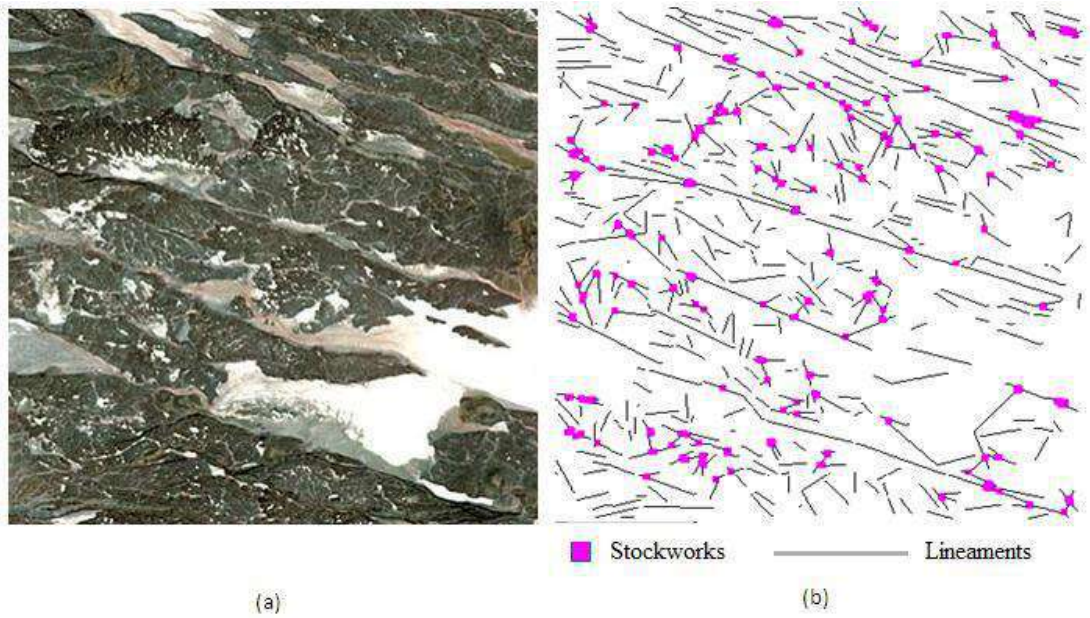


Fig. 4: (a) Original input image of afar region (b) Lineaments & stockworks extraction using LIST algorithm.

Table 1: Geospatial analysis – lineaments length analysis

Methods	Micro lineaments	Moderate lineaments	Macro lineaments	Total no. of lineaments
MBC	106	7	17	130
FCM	224	143	278	645
LIST	391	107	201	699

Table 2: Geospatial analysis – lineaments length classification and its contribution (in percentage)

Methods	Micro lineaments	Moderate lineaments	Macro lineaments
MBC	81.53%	5.38%	13.07%
FCM	34.72%	22.17%	43.10%
LIST	55.93%	15.30%	28.75%

Table 3: Geospatial analysis – Lineaments length classification and its total length

Methods	Micro lineaments	Moderate lineaments	Macro lineaments	Total length of lineaments
MBC	26.51	9.68	269.86	306.05
FCM	121.62	206.02	6902.67	7230.31
LIST	162.66	158.997	17240.33	17561.99

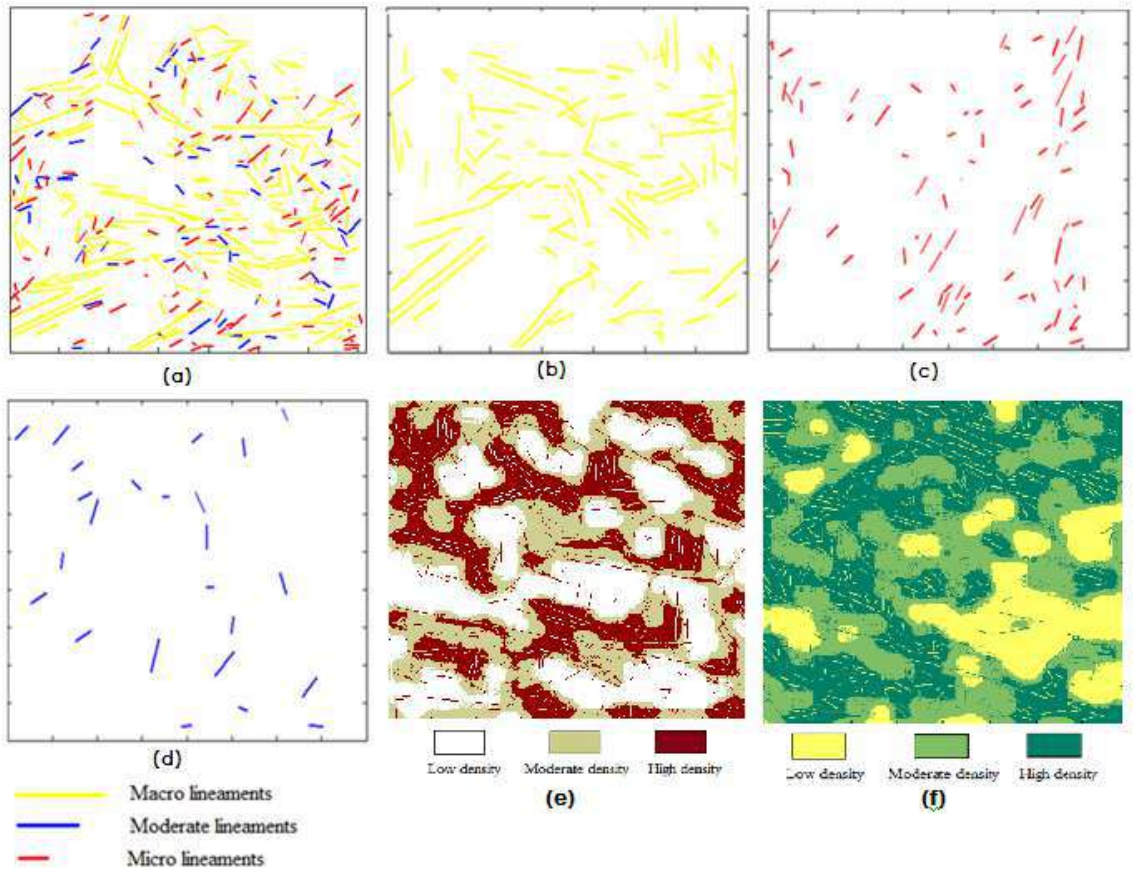


Fig. 5: (a) Geospatial analysis by proposed LIST algorithm. Length analysis: (b) Macro lineaments (c) Micro lineaments (d) Moderate lineaments. Density analysis: (e) Stockworks density (f) Lineaments density

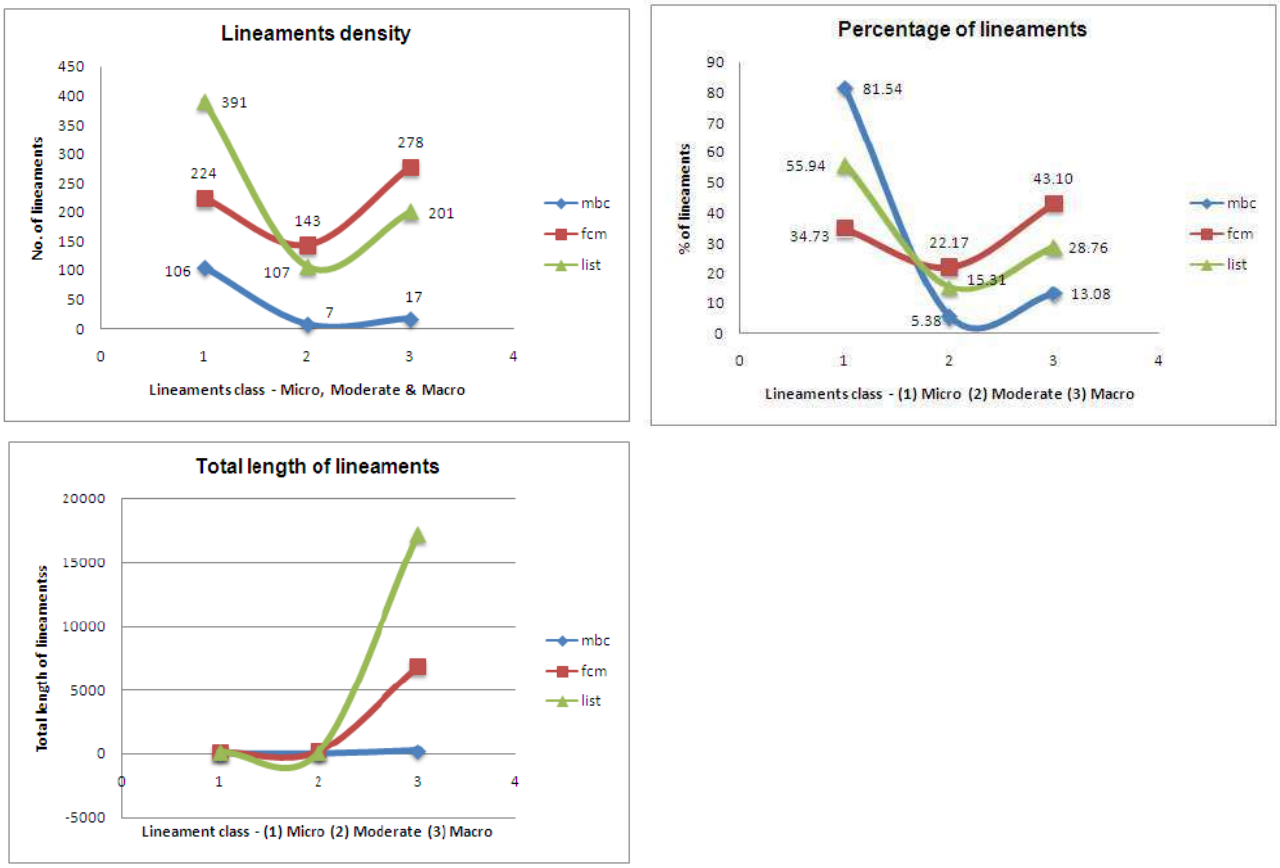


Fig. 6: Lineament density analysis and length analysis

IIOAB JOURNAL
ISSN: 0976-3104

A total of 699 lineaments are mapped in the afar depression region and its surroundings using the proposed LIST algorithm – a manual extraction technique. The total length of lineament is 17561kms. From the visual interpretation, all the lineaments mapped in the region have its orientation towards NE-SW direction. This detailed analysis of lineaments length is consistent with the published geological survey map of Ethiopia. Lineament density analysis is also done to calculate the frequency of lineaments based on its length. Micro lineaments extracted using LIST algorithm contributes 55% of overall lineaments. Mask based clustering (MBC) and Fuzzy C Means (FCM) algorithm contributes 81% and 34%. These two methods are applied in afar depression image results 130 and 645 lineaments with a total length of 306 and 7230kms respectively. This difference is caused by different methods of mapping on regional and local scale and also some of the lineaments are missing due to under segmentation and over segmentation problem. In this work, stockworks density map is also generated to strengthen the lineament density analysis. Areas with higher micro lineaments density and higher stockworks density are the main prospective zones for mineralization. In conclusion, the geospatial analysis result shows that the identification of geological lineaments over wide area is possible and detailed information for the purpose of exploring minerals and its prospective zones can be obtained using the proposed LIST algorithm.

CONCLUSION

Remote sensing (Image processing) and GIS (modelling) techniques is very much helpful to map the geologic lineaments. In particular, the flow of minerals in hard formations is mainly controlled by the lineaments corresponds to the faults, fractures and stockworks. This work concentrates on extracting both the lineaments and stockworks. The complexity lies while detecting the lineaments due to poorly illuminated image, global scale segmentation, and discontinuity in lineament patterns. In this work, all these problems are solved effectively. CIE L*a*b* color space and Brightness Preserving Dynamic Fuzzy Histogram Equalisation technique eliminates the poor illumination effect and enhances the image for further processing. Then the image is splitted into equal sized grids to locate micro lineaments. 24-element matrix and 8-element matrix are used to detect line segments by considering its contribution of neighbour pixel values; to avoid discontinuity in linear patterns using the proposed LIST algorithm. It also targets the stockworks easily. This gives better results when compared to the existing methods of lineament segmentation. Lineament density map and stockworks density map are also generated from this work.

FUTURE ENHANCEMENTS

Mineralogical application of remote sensing helps to identify the large scale lineaments, which in turn helps to understand the geologic structures / topologies of an area. Different combination of sensors mounted on the satellite / aircraft is used to facilitate accurate mapping of lineaments. The research work will be enhanced for the following applications in future. Modernization and Globalisation are some of the cause for the occurrence of natural hazards. Many events such as earthquakes, landslides, floods can be dangerous to people. Geologists may predict these events well in advance to avoid damages to the property and life of the people, with the help of the geological map / lineaments map prepared with the knowledge of past activity and earth's surface. This effort will help in the hazards susceptibility analysis. Vulnerability mapping of these areas will increase the Resilience of Communities.

The current framework can be extended to deal with the factors of geology, such as tectonics, geophysics, geochemistry to map, interpret and classify the minerals from the prospective zones. The knowledge of lineaments can be applied to those terrain surfaces to map/monitor the road networks, drainage density networks, alignment of vegetation, etc. and the effect of the change in these networks. A more precise determination of ground water is mainly controlled by the low permeability areas such as rock types, landforms, geological structures, soil, land use etc. hydro geological and geophysical methods are used to generate the groundwater prospect zone which is essential to delineate groundwater reservoir.

CONFLICT OF INTEREST
The author(s) declare(s) that there is no conflict of interest regarding the publication of this article

ACKNOWLEDGEMENTS
Authors are grateful to NASA's Earth Observatory., USA who offers the Afar depression open source ASTER sensor image for this research work. The authors thank the anonymous reviewers for their helpful comments that significantly improved the manuscript.

FINANCIAL DISCLOSURE
The author declares that he has no relevant or material financial interests that relate to the research described in this paper.

IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] Hobbs WH.[1912] Earth Features and Their Meaning: An Introduction to Geology for the Student and General Reader", Macmillan, New York, NY (347pp),

[2] Gurugnanam Suresh, Kalaivanan,[2014] Extraction of Lineament And Lineament Density Assessment From Satellite Data In Kolli Hill, Tamil Nadu, South India, International Journal of Recent Scientific Research 5 (7):1365-1367

[3] Raj NJ, Prabhakaran A, Muthukrishnan A. [2017] Extraction and analysis of geological lineaments of Kolli hills, Tamil Nadu: a study using remote sensing and GIS. Arabian Journal of Geosciences, 10(8): 195.

[4] Rayan Ghazi Thannoun.[2013] Automatic Extraction and Geospatial Analysis of Lineaments and their Tectonic Significance in some areas of Northern Iraq using Remote Sensing Techniques and GIS, International Journal Of Enhanced Research In Science Technology & Engineering, 2(2)

[5] Dasho OA, Ariyibi EA, Akinluyi FO, Awoyemi MO, Adebayo, AS. [2017] Application of satellite remote sensing to groundwater potential modeling in Ejigbo area, Southwestern Nigeria. Modeling Earth Systems and Environment, 1-19.

[6] Acharya D, Porwal A, Bhattacharya A. [2017]. Remote detection of geological structures: an application to the Aravalli region, western India. Geocarto International, 32(3), 257-273.

[7] Hung LQ, Batelaan O.[2003] Environmental geological remote sensing and GIS analysis of tropical karst areas in Vietnam, Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGAR SS), Toulouse, France, 21-25 July 2003, 4:. 2964-2966.

[8] Kim GB, Lee JY, Lee KK. [2004] Construction of lineament maps related to groundwater occurrence with arcview and Avenues scripts, Computers &Geosciences, Vol. 30, 1117–1126.

[9] Costa RD, Starkey J. [2001] Photo Lin: a program to identify and analyze linear structures in aerial photographs, satellite images and maps, Computers and Geosciences, 27(5): 527-534.

[10] Spatial density estimation based segmentation of super-resolution localization microscopy images Kuan-Chieh, Jackie Chen, Ge Yang, Jelena Kovacevic, Proceedings of the IEEE International Conference on Image Processing (ICIP), 2014, 867-871.

[11] Chan AB, Z SJ Liang, Vasconcelos N. [2008] Privacy preserving crowd monitoring: Counting people without people models or tracking. CVPR.

[12] Ryan D, Denman S, Fookes C, Sridharan S. [2009] Crowd counting using multiple local features. DICTA '09: Proceedings of the 2009 Digital Image Computing: Techniques and Applications, pp. 81–88,

[13] Kumar NS, Arun, M. [2016] Accuracy analysis of various classification algorithms for used land. International Journal of Enterprise Network Management, 7(2): 113-132.

[14] Kumar NS, Arun M. [2015] Enhanced classification algorithms for the satellite image processing. Indian Journal of Science and Technology, 8(15).

[15] Zeeb C, Gomez-Rivas E, Bons PD, Virgo S, Blum P.[2013] Fracture network evaluation program (franep): a software for analyzing 2D fracture trace-line maps, Computers and Geosciences, 60: 11–22.

[16] Cheng YB, Mao JW, Chang ZS, Pirajno F.[2013] The origin of the world class tin-polymetallic deposits in the Gejiu district, SW China: constraints from metal zoning characteristics and 40Ar–39Ar geochronology," oregeology Reviews, 53: 50–62

[17] Zhao J, Zuo R, Chen S, Kreuzer OP.[2015] Application of the tectono-geochemistry method to mineral prospectivity mapping: a case study of the Gaosong tin-polymetallic deposit, Gejiu district, SW China, Ore Geology Reviews, 71: 719–734,

[18] Khosroshahizadeh S, Pourkermani M, Almasian M, Arian, M, Khakzad A. [2016] Lineament Patterns and Mineralization Related to Alteration Zone by Using ASAR-ASTER Imagery in Hize Jan-Sharaf Abad Au-Ag Epithermal Mineralized Zone (East Azarbaijan—NW Iran). Open Journal of Geology, 6(04): 232.

[19] Kusák M, Krbcová K, Križan F, Kunc J, Bilková K, Barlík P., Hotový O. [2017] Analysis of the relationship of automatically and manually extracted lineaments from dem and geologically mapped tectonic faults around the main ethiopian rift and the ethiopian highlands, ethiopia. Development, 5:17.

[20] Pothiraj Prabu and Baskaran Rajagopalan. [2013] Mapping of Lineaments for Groundwater Targeting and Sustainable Water Resource Management in Hard Rock Hydrogeological Environment Using RS- GIS, Climate Change and Regional/Local Responses, Dr Pallav Ray (Ed.), intech, DOI: 10.5772/55702.

[21] Hhttp://www.see.leeds.ac.uk/afar/new-afar/geology afar/structure-tech-pages/geol-afar-dep-tech.html (Geology of Afar depression – University of Leeds)

[22] http://asterweb.jpl.nasa.gov/gallery-detail.asp?Name=afar

[23] https://en.wikipedia.org/wiki/Afar_Region (Afar Region - Wikipedia)

[24] Mustafa Ali Hassan, Safaa Sabah Adhab,[2014] "Lineament automatic extraction analysis for Galal Badra river basin using Landsat 8 satellite image", Remote Sensing Department, College of Sciences, University of Baghdad, 12(25): 44-55, Iraqi Journal of Physics

[25] Sukumar M. [2016] Detecting linear structures within the ASTER satellite image by effective denoising and contrast enhancement in the device independent color space. Advances in Natural and Applied Sciences, 10(10 SE): 18-24.

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

REVIEW

REVIEW ON SENTIMENT ANALYSIS A LEARNERS' OPINION

Jayakumar Sadhasivam[1*], Ramesh Babu Kalivaradhan[2]

[1]School of Information Technology and Engineering, VIT University, Vellore – 632014, TN, INDIA
[2]School of Computer Sciences and Engineering[2], VIT University, Vellore – 632014, TN, INDIA

ABSTRACT

With the booming growth of learning in the 21st century and it is keep expanding it is branches towards many fields. Many researchers expressed their ideas in education data mining filed using opinion mining also called as Sentiment Analysis (SA). Education data mining is used to analyze the learner, instructor and provide the better way to the learners. Sentiment analysis uses Natural Language Processing (NLP) text analysis and computational linguistics to detect and extract our emotions, opinions, attitude and content extracted from individuals using text mining techniques. In this survey paper, deep focused on sentiment analysis in the field of education and learning. The ultimate aim of this paper is to represent the current Sentiment Analysis (SA) classification, models, algorithm, and applications.

KEY WORDS
Sentiment Analysis (SA), learning, Data Mining, Natural Language Processing (NLP), Education

Received: 18 May 2017
Accepted: 28 July 2017
Published: 20 Sept 2017

***Corresponding Author**
jayakumars@vit.ac.in

INTRODUCTION

Over a decade, learning is revolutionized using information and communication technologies. Learning platform and web2.0 technologies improved the way of learning and teaching. By increasing the use of the Internet for learning, discussion forums and learning related activities, a huge amount of unstructured data from the learners. In this digital era, learner's discussion forum, comments, and reviews are very important for the instructor. It made difficult to Instructors to understand the learner's mindset and feedback. The amount of data is huge to analyze. Such data can be analyzed using different mining techniques. Focusing on data and text mining techniques to enhance learner's ability is emphasized. Information is related to user feedbacks, discussion forum content, comments and reviews are utilized to the analysis.

In this survey paper, we are discussing research articles related to sentiment analysis in the educational/learning domain. Survey revealed many researchers used the different type of sentimental models and algorithms, but most of the researchers focused on Naive Bayes (NB) and Support Vector Machine (SVM) algorithm. Weka[1] tools are the most common tool used by researchers for analyzing the sentimental analysis.

SA uses the NLP, Text Analysis and Computational Linguistics to extract emotions and feeling from the raw data. SA is an example where a combination of knowledge from engineering, statistics and linguistics is essential for providing accurate analysis. Researchers got attention on SA due to its potential application in information retrieval, discussion forum, blogs, reviews by helping them to retrieve the user opinions from their content.

An essential task in opinion minion is to categorize the polarity of a given content in the document, sentence, or feature or aspect level — whether the given content opinion is a document, a sentence or an entity feature/aspect to be positive, negative, or neutral[2]. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy"[2]. Sentiment polarity in certain contexts needs in order to predict it from the user-generated text data. Information data can be divided into two categories: facts and opinions. Facts are the objective expression of their entities, events and their properties. Perspectives are typically subjective expressions that portray individuals' sentiments, examinations or emotions toward entities, events and their properties.

With respect to educational domain, SA is included in applying the automatic content analysis process for the purpose of removing the unrelated words and gathering the words related to sentiment from the user opinions and also indicate the variety of sentiment conveyed in education domain like learning website, blogs, discussion forums and review sites. Instructor or educational website administrator tracks the user opinions to find the difficulties of learning from the educational platform. It helps to improve the educational strategies and provide the benefits to the instructor. SA plays an important role in developing a better learning platform and provide to the learner in a better way and also a new user to enroll in their learning platform.
As an emerging area, millions of students and learners enrolling and signup for the new course in MOOC platform (Coursera, edX, Udacity etc..). Currently, most of the university offer the course, lecture notes, video lecture, assignment and text via Learning Management System (LMS). Learning platform is growing each and every day and providing newer and better opportunities for students to learn via the Internet and other Information and Communication Technology. In the educational domain, it is necessary in order to the course provider or instructor to provide the communication between each and every user via discussion board, chat rooms or comments to provide their opinions and evaluate their work. The fundamental need of LMS is to provide the learners requirement very effectively. LMS crucial task is to get

IIOAB JOURNAL
ISSN: 0976-3104

the user opinions via survey or comments to know, whether they are providing the user needs efficiently as possible. This survey assists in LMS administrator to track the learner need and rectify their problem immediately or in the upcoming update.

This paper is organized into various sections. Section 2 with literature review, Section 3 with Data Source for Sentiment Analysis in Education Domain, Section 4 for Sentiment Analysis Table summary of the review and finally Section 5 concludes the survey.

LITERATURE REVIEW

At the early stage of sentiment analysis is processed based on the thumbs up and thumbs down. Peter D. Turney[3]used this method to find out the sentiment analysis in Epinions websites, now acquired by eBay.

Opinion mining is the primary part of Sentiment Analysis. The analysis content is from the review, blog, discussion forum, comments etc. The analysis is used to identify the sentiment classification in the content is generally based on positive and negative. A content has a positive semantic orientation when it has good associations (e.g. “subtle nuances”) and a negative semantic orientation when it has bad associations (e.g., “very cavalier”)[3][4]. Sentiment classification findings which achieve to conclude whether a text is objective or subjective, or whether a subjective text holds positive or negative sentiments[5]. Sentiment classification uses different types of machine learning and natural language processing approach. There are many supervised and unsupervised techniques available for sentiment classification. Support Vector Machine (SVM) and Naive Bayes(NB) are supervised learning techniques. K-means clustering is used for unsupervised learning techniques. Tony Mullen and Nigel Collier[6] used the SVM to find out the sentiment Analysis in the Internet Movie Database using the Hybrid Support Vector Machine with combine unigram-style feature based SVM.

Vishal A. Kharde and S.S. Sonawane[7] used the Naive Bayes techniques and much more to find out the sentiment analysis in the twitter tweets unstructured data. Sentiment Analysis can be better predicted from the minimum words in the real-time environment using the twitter social media. It helps to identify the group of tweets related to a specific domain using the tags. M. Di Capua, E. Di Nardo, and A. Petrosino[8] proposed an architecture to identify the e-learning tags in the twitter data to identify the SA using machine learning components. As e-learning become the emerging platform in the educational domain, researchers focused their research towards the e-learning and sentimental analysis. D. Song, H. Lin, and Z. Yang[9] proposed opinions and adopting automatic opinion mining to recognize the sentiment of opinions from the web pages on which users are considering or relating their personal opinions and evaluation of the services. Understanding this process of emotional and reaction of a student is complicated in the learning environment. H. H. Binali, C. Wu, and V. Potdar[10] used the conceptual emotion detection and analysis in the system for education domain using the sentimental analysis techniques in education blogs and reviews.

As we are located in the online world, everything is accessed and purchased via online, the online market with trillions of markets share worldwide. D. Tayal and S. Komaragiri[11] proposed their sentimental analysis review towards the blog or microblog, they expressed an idea how a post may change the product sales up and down and to predict the stock market performance. A sample data of 520 movie reviews are used to analyze the opinion mining, the movies are based on particular genres like blog posting, discussion threads, user reviews and critical review. The analysis is tremendously engrossed on the document, sentence length, part-of-speech(POS) distribution, vocabulary, aspects of movies discussed, star ratings used and multimedia content in the reviews. This study identifies the positive and negative on different genres[12]. Baojun Ma[13] proposed their sentimental analysis using the clustering techniques based on the online review. Researchers used the vector space modeling(VSM) k-means clustering algorithm to perform the sentimental analysis to identify the positive and negative review of Movie and Travel website online reviews as a dataset. SA now ahead forward towards analyzing the sentimental in the video, audio, and visual clips. Previously opinion mining analysis is carried out in the textual content. Moise´s H. R. Pereira and his team[14] used the facial recognition techniques to identify the sentimental analysis in the news video. They compute attributes, such as visual potency of recognized feelings and emotions, field sizes of users, voicing probability, sound loudness, speech fundamental frequencies and the sentiment outcomes (polarities) from text sentences in the closed caption[14]. A dataset of 520 videos from Brazilian and American TV newscasts used for SA. After the SA approach, they identified that an accuracy of up to 84% in the sentiments (tension levels) classification task, thus demonstrating its high potential to take advantage of media analysts in several applications, especially, in the journalistic domain[14].

In the current digital world, everyone deals with the problem of email spam threads, affecting millions of users per day. Enaitz Ezpeleta[15] and his team developed a spam classifier using the Bayesian spam filtering classifier. That can improve the spam filtering classifier adding the polarity message. SA utilized to detect the spam emails and they achieved the level of 99.21% accuracy in spam classification. Currently,

IIOAB JOURNAL ISSN: 0976-3104

social media are median or as a tool used by industry to attract the customer. The industry is watching their social media carefully to know their current updates by their consumer. A single tweet or Facebook post, can affect the industry or company in the larger manner. Social media provides all the content public access and share with other users. It assists industry into gain or loses the market based on the user post. Wu He[16] used the social media tweets to analyze the competitive pizza industry in china. They used unstructured data from the Facebook post, and Twitter tweets of Pizza Hut, Domino's and Papa John's Pizza. This social media tweet is utilized to get the market and e-market decision in china about the pizza industry. Analyzing the content always gives some unpredictable things, that helps the industry for their growth, SA is proposed and analyzed in different categories. We found unbelievable information and answer from the public forums and discussion threads. Yahoo Answer constitutes one of the community-driven questions and answers forum. Onur Kucuktunc[17] proposed a way to analyze the yahoo answers and extracted the user information via sentimental analysis. They used the yahoo answers to identify the user's best answers and sentiment based answers in each domain.The authors[18] introduced a particular structured framework that takes an advantage of existing MOOCs data to provide a learning style for MOOC user. By recognizing and understanding the learning styles, one can utilize certain procedures and approaches to enhance the rate and nature of learning.

DATA SOURCE

Websites and blogs

With an increasing using of Internet, learning websites and blogs are mostly utilized to express the user opinions, feeling, recommendation etc. towards anything. Blogs and microblogs are used commonly to express the daily events happening towards the learners. Learning is expressing their feeling, emotions, opinions as positive and negative in a different way publicly. These feelings, emotions, and opinions are utilized to analyze the learner's perspective towards a particular course or subject. Dan Song, Hongfei Lin, Zhihao Yang[9] analyzed 446 learning review articles in Chinese language using SVM. H. H. Binali, C. Wu, and V. Potdar[10] analyzed e-learning corpus of students weekly online review for a semester towards a course. Z. Kechaou, M. Ben Ammar, A. M. Alimi and M. Ammar[19] gathered data from 5 different e-learning blogs reviews and Moodle forum data to analyze 1000 positive and 1000 negative text emotions of the user to classify the sentimental classification model. F. Tian, H. Liang, L. Li, and Q. Zheng[20], analyzed 9957 data corpus from Xi'an Jiaotong University community services. The researchers manually labeled the content into three emotions types, positive, negative and neutral in the Chinese language.

Discussion forums

Recent trends are to discuss anything in the discussion forums, previously it was only used as a question and answer forum for the user, but nowadays the discussion forums are used for anything to discuss privately and publicly. Discussion forum affects the learners in a different way to show their opinion. Reviews, recommendations etc. Alaa El-Halees[21] extracted discussion forum data of 22MB contains 4957 discussion posts of five courses in the Arabic language.

Feedback systems

Each and every organization, company and institution use feedback system get the user opinions, comments, and reviews as a known user or anonymously. These feedbacks are used in a variety of ways to analyze the data and improve the institutions, company or organizations. P. Šaloun, M. Hruzík, and I. Zelinka[22] used the student's feedback using adaptive web system XAPOS to get the data from the e-shop and e-learning domain in Slavic and Czech language. N. Altrabsheh, M. M. Gaber and M. Cocea[23] used the student feedback system via clickers, mobile phones, clicker applications and social media data for their analysis.

Summary of the Review

Table 1: Study of sentiment analysis in the learning domain

S.no	Studies	Language	Mining Technique	Description	Data Source
1.	Dan Song, Hongfei Lin and Zhihao Yang (2007)[9]	Chinese	Support Vector Machines	Learning system identifies the drawback from the user opinion on the course materials, teachers feedback reviews and comments.	Review dataset of 446 articles (7,324 sentences) are chosen from the different learning domain.
2.	Haji H. BINALI, Chen WU, Vidyasagar POTDAR (2009) [10]	English	Naïve Bayes	Appraisal theory, corpus sentence and statistical schema matching are used to identify the emotions in the student submitted weekly review[10].	Student posted weekly online review for a semester for a course
3.	Alaa El-Halees (2011) [21]	Arabic	Naive Bayes	Student submit their review about the course, it used to measure the course performance and compare with one or more course in each lecture or semester.	4,957 discussion posts which contain 22 MB. 167 Posts, 5017 statements, Total No. of words 27456.
4.	Zied Kechaou, Mohamed Ben	English	Naïve Bayes and Support	Reviews are gathered from a multitude of e-learning blogs and	E-learning review of 1000 positive and 1000 negative

	Ammar, Adel.M Alimi (2011)[19]		Vector Machine	an analysis to study the nature and the structure of web discussion forums and learning blogs turns out to be a significant endeavor[19].	corpus[19].
5.	A. Nisha Jebaseeli, Dr. E. Kirubakaran (2012) [24]	English	Naïve Bayes	Analysis of the M-learning system using Naive Bayes algorithm and compared with K-nearest neighbor and random forest data mining algorithm.	100 reviews are selected in each category of positive, negative and neutral from online android market place.
6.	Feng Tian, Huijun Liang, Longzhuang Li, Qinghua Zheng (2012)[20]	Chinese	Support Vector Machine and Naïve Bayes	Interactive Chinese Texts is used for text classification later combined with syntax feature sets, two sets of new features, frequency based features and interaction related features, which are different from the traditional feature set. Feature set is performing better than SVM[20].	Data collected from student community service includes 9957 turns manually labeled with three emotion types, positive, negative and neutral[20].
7.	P. Šaloun, M. Hruzík, I. Zelinka (2013) [22]	Slavic and Czech	Support Vector Machine	Analysis of user feedback in the Slavic and Czech language text content based on business and education environment. The contents are automated for computer processing to check the positive and negative sentiment from the text.	Data collected from Student's feedback using adaptive web system XAPOS.
8.	Nabeela Altrabsheh, Mohamed Medhat Gaber, Mihaela Cocea (2013)[23]	English	Naive Bayes and Support Vector Machines	Student Response Systems used to collect the feedback for lecturer via social media using clickers and mobiles.	Data source from clickers, Mobile Phones (Clicker Applications, SMS), Social Media.
9.	Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, Jibao Gu (2014) [25]	English	Naive Bayes and Support Vector Machine	Performance comparative assessment for ensemble methods (Bagging, Boosting, and Random Subspace) based on five base learners' algorithm such as Naive Bayes(NB), Maximum Entropy, Decision Tree, K Nearest Neighbor, and Support Vector Machine(SVM)[26]. Ten public SA datasets were investigated to confirm the effectiveness of ensemble learning. To increase the functionality of individual base learners for sentiment classification[25].	Dataset from MPQA, Movie review, Two web forums, NTCIR Opinion, Product review and Chinese review. Total of 1200.
10.	Balaji Jagtap, Virendrakumar Dhotre (2014)[27]	English	Support Vector Machine	Teacher feedback system automates the student's feedback for a particular teacher to provide positive and negative, the analysis performed on different domains. On applying advanced feature selection method with the hybrid approach of sentiment classification. Hybrid approach works well with complex data.	Data is collected from student in unstructured format
11.	Miaomiao Wen, Diyi Yang, Carolyn Penstein Rose (2014)[28]	English	Sentiment Polarity Analysis	MOOC forums are used to identify the trending opinions towards the course. To observe the students, drop out and each day forum post. Methods used - Course-level Sentiment Analysis: Collective Sentiment Analysis, User-level Sentiment Analysis: Survival Analysis.	Dataset from three courses from coursera
12.	Lorenzo A. Rossi, Omprakash Gnawali (2014)[29]	English, French, Spanish and Chinese	Support Vector Machine	Analysis of four coursera courses discussion threads. To Analyze the forum activities, different languages, interaction among other users, study groups, assignment and lectures[29].	Discussion forums of 60 Coursera (99,624 threads, 739,093 posts and comments), downloaded between August 2013 and April 2014[29].
13.	P.Bharathisindhu, S. Selva Brunda (2014)[30]	English	Naive Bayes and Support Vector Machine	Identifying the sentimental analysis in E-portal based on the user interest or area, information extracted from subjective information, recognizing opinion-oriented questions and summarization.	Data collection of 100 users review from the website Functionspace.org.

14.	Aysu Ezen-Can, Kristy Elizabeth Boyer, Shaun Kellogg, Sherry Booth (2015)[31]	English	Bayesian Information Criterion (Bic)	An unsupervised dialogue act classification framework with MOOC modeling approaches, with the primary goal of gaining insights about the structure of forum posts in an MOOC[31].	Data collected within an 8-week MOOC for Educators (MOOC-Ed) titled Planning for the Digital Learning Transition in K-12 Schools[31].
15.	V. Kagklis, A. Karatrantou, M. Tantoula, C. T. Panagiotakopoulos, and V. S. Verykios (2015)[32]	Greek	Dictionary-Based Sentiment Classification.	Post graduate students online course forum used to analyze the Sentiment, text mining and Social Network Analysis(SNA) techniques from the data. To analyze the course performance and interaction among students. Sentiment analysis is used in different levels such as sentence, document, and corpus level[33].	Data set consists of the forum activity of 64 students. 12% of women and 88% of Men. A total of 371 messages were posted. 89 out of 371 were starting posts, while the rest 282 of them were replies. 198 messages were posted by the students and the rest were posted by tutors
16	Devendra Singh Chaplot, Eunhee Rhim, Jihie Kim (2015)[34]	English	Sentiwordnet	Coursera Forum posts are used for sentimental analysis to find out the student attrition and study effectiveness using neural network modelling.	Data source from Coursera forum over 3 million students active click logs and over 5000 forum posts are used for research work[34].
17	S. Priyanka, M. Sivakumar (2015)[35]	English	Support Vector Machine	Facebook comments are used for social data analysis based on context adaptive system[35].	KONECT (Koblenz Network Collection) students' textual feedback: SVM Algorithm used (with three types of kernel). A dataset of 1036 instances of teaching and learning related feedback was used, which was analyzed and categorized by 3 experts[35].
18	T. Zarra, R. Chiheb, R. Faizi, and A. El Afia (2016)[36]	English	Latent Semantic Analysis, Singular Value Decomposition	To identify the students' opinions about educational issues that are problematic	62505 Stackoverflow comments

CONCLUSION

The most recent decade has seen tangible advances in the areas of Natural Language Processing(NLP) and Semantic Analysis(SA) or Opinion Mining. Intellectual developments and its expanded computational power have achieved applications that distinguish subjects and sentiments in correspondences, consequently group of unstructured data information in enterprise settings. This paper reviews the methods applied to various problems in education and learning and it also portrayed the need for the sentiment analysis in the field of education and learning. Most of the researchers used Naïve Bayes and Support Vector Machine algorithm for their research work because of its accuracy and ease of implementation. Furthermore, reviewers used data sources focusing not only on the English language, but also on their native language for sentimental analysis. The future work will suggest a comprehensive model and a case study on the impact of various models in a real-time scenario.

CONFLICT OF INTEREST
NONE

ACKNOWLEDGEMENTS
NONE

FINANCIAL DISCLOSURE
NONE

REFERENCES

[1] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. [2009] The WEKA data mining software,"SIGKDD Explor. Newsl. 11(1): 10.

[2] Sentiment Analysis,"wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Sentiment_analysis.

[3] Turney PD.[2002] Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews, Proc. 40th Annu. Meet. Assoc. Comput. Linguist., 417–424.

[4] Matt Kiser, Introduction to Sentiment Analysis Algorithms, algorithmia.com, 20-Jan-2017. [Online]. Available: https://blog.algorithmia.com/introduction-sentiment-analysis-algorithms/.

[5] Abbasi A, Chen H, Salem A.[2008] Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums,ACM Trans. Inf. Syst. ..., 26(3): 1–34.

[6] Mullen T, Collier N.[2004] Sentiment Analysis using Support Vector Machines with Diverse Information Sources,"Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. (EMNLP 2004), pp. 412–418.

[7] Kharde VA, Sonawane SS.[2016] Sentiment Analysis of Twitter Data: A Survey of Techniques, Int J Comput. Appl., 139(11): 975–8887.

[8] Di Capua M, Di Nardo E, Petrosino A. [2007] An Architecture for Sentiment Analysis in Twitter, 214–221.
[9] Song D, Lin H, Yang Z. [2007] Opinion mining in e-learning system,Proc. - 2007 IFIP Int. Conf Netw Parallel Comput Work. NPC 2007, pp. 788–792.
[10] Binali HH, Wu C, Potdar V. [2009] A new significant area: Emotion detection in E-learning using opinion mining techniques,2009 3rd IEEE Int. Conf. Digit. Ecosyst. Technol. DEST '09, pp. 259–264.
[11] Tayal D, Komaragiri S.[2009] Comparative analysis of the impact of blogging and micro-blogging on market performance, Int J Comput Sci Eng. 1(3): 176–182.
[12] Na JC, Thet TT, Khoo CSG. [2010] Comparing sentiment expression in movie reviews from four online genres,"Online Inf Rev, 34: 317–338
[13] Ma B, Yuan H, Wu Y. [2015] Exploring Performance of Clustering Methods on Document Sentiment Analysis,
[14] Pereira MHR, Pádua FLC, Pereira ACM, Benevenuto F, Dalip DH. [2016] Fusing audio, textual, and visual features for sentiment analysis of news videos,Proc. 10th Int. Conf. Web Soc. Media, ICWSM 2016, no. Icwsm, pp. 659–662.
[15] Ezpeleta E, Zurutuza U.[2016] Does sentiment analysis help in bayesian spam filtering ?.
[16] He W, Zha S, Li L.[2013] Social media competitive analysis and text mining: A case study in the pizza industry, 33: 464–472.
[17] Kucuktunc O, Weber I, Cambazoglu BB. [2012] A Large-Categories and Subject Descriptors, 633–642.
[18] Sadhasivam J, Babu R.[2017] MOOC : A FRAMEWORK FOR LEARNERS USING LEARNING STYLE, Int. Educ. Res. J. 3(2): 21–24.
[19] Kechaou Z, Ben Ammar M, Alimi AM, Ammar M, Improving e-learning with sentiment analysis of users' opinions, IEEE Glob. Eng. Educ. Conf. EDUCON [2011], 1032–1038.
[20] Tian F, Liang H, Li L, Zheng Q. [2012] Sentiment classification in turn-level interactive Chinese texts of e-learning applications,"Proc. 12th IEEE Int Conf Adv Learn Technol. ICALT 2012, pp. 480–484.
[21] El-Halees A.[2011] Mining Feature-opinion in Educational Data for Course Improvement, Int J New Comput. Archit. Their Appl. 1(4): 1076.
[22] Šaloun P, Hruzík M, Zelinka I. [2013] Sentiment analysis - E-bussines and E-learning common issue, ICETA 2013 - 11th IEEE Int. Conf. Emerg. eLearning Technol. Appl Proc, 339–343.
[23] Altrabsheh N, Gaber MM, Cocea M.[2013] SA-E: Sentiment analysis for education,"Front. Artif. Intell. Appl., 255: 353–362
[24] Jebaseeli AN. [2012] M-Learning Sentiment Analysis with Data Mining Techniques, Int J Comput Sci. 3(8): 45–48.
[25] Wang G, Sun J, Ma J, Xu K, Gu J. [2014] Sentiment classification: The contribution of ensemble learning,"Decis. Support Syst., 57(1): 77–93
[26] Wang G, Zhang Z, Sun J, Yang S, Larson CA,[2015] POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis,"Inf. Process. Manag. 51(4): 458–479
[27] Jagtap B, Dhotre V,[2014] SVM and HMM Based Hybrid Approach of Sentiment Analysis for Teacher Feedback Assessment,Ijettcs.Org, 3(3): 229–232.
[28] Wen M, Yang D, Rosé C,[2014] Sentiment Analysis in MOOC Discussion Forums: What does it tell us?,Proc. Educ. Data Min., no. Edm, 1–8
[29] Rossi LA. and Gnawali O . [2014] Language independent analysis and classification of discussion threads in Coursera MOOC forums," in Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on, 2014, pp. 654–661.
[30] P. Bharathisindhu and Brunda SS.[2014] Identifying E-Learner's Opinion Using Automated Sentiment Analysis in E-Learning,IJRETInternational J Res Eng Technol. 3(1): 2319–2322
[31] Ezen-Can A, Boyer KE, Kellogg S, Booth S.[2015] Unsupervised modeling for understanding MOOC discussion forums,"Proc. Fifth Int. Conf. Learn. Anal. Knowl. - LAK '15, pp. 146–150,
[32] Kagklis V, Karatrantou A, Tantoula M, Panagiotakopoulos CT, Verykios VS. [2015] A Learning Analytics Methodology for Detecting Sentiment in Student Fora: A Case Study in Distance Education.,"Eur. J. Open, Distance E-Learning, 18(2): 74–94
[33] Agathangelou P, Katakis I, Kokkoras F, Ntonas K. [2014] Mining Domain-Specific Dictionaries of Opinion Words, in Web Information Systems Engineering -- WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part I, B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, and Y. Zhang, Eds. Cham: Springer International Publishing, 2014, pp. 47–62.
[34] Chaplot DS, Rhim E, Kim J. [2015]Predicting student attrition in MOOCs using sentiment analysis and neural networks, Work. 17th Int. Conf. Artif Intell Educ AIED-WS, 1432: 7–12.
[35] Priyanka S, Sivakumar M.[2015] Big Data Processing in Sentiment and Opinion Mining for Detecting Student Depression in E-Learning Using Rich Facebook Dataset Collection, 5(4): 1208–1216.
[36] Zarra T, Chiheb R, Faizi R, El Afia A.[2016] Using Textual Similarity and Sentiment Analysis i n Discussions Forums to Enhance Learning.pdf, 10(1): 191–200

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

AN EFFICIENT HYBRID CRYPTOGRAPHY FOR SECURED DATA IN PUBLIC CLOUD ENVIRONMENT

Prabu S*, Gopinath Ganapathy

Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappalli, Tamil Nadu, INDIA

ABSTRACT

Background: *Cloud Computing (CC) can be termed as the figuring which is based on Internet in which powerful apportioned servers are giving programming's including diverse belonging with encouraging for purchasers on a compensation as-you-utilize hypothesis. The archive in cloud is nothing but securing data on outsider cloud servers. The reasons for CC being huge are boundless capacity, reinforcement along with restoration. The CC's bad marks are specific concern, value including nonattendance of help. Regardless to say, reliability is the essential hitch. Since the capacity of data on outsider cloud servers' providers is concerned, one cannot say that it is fully secured. This constrains a sudden risk. Various servers in the cloud are concerned which means that they endeavor in order to examine the data secured on it.* ***Methods:*** *The paper is going to deliver software which helps to upgrade the reliability of cloud uses parcel along with an encoding approach that can intensify insurance of the cloud. Firstly, getting the purchaser's affirmation and then parceling that to various segments has been finished. Once the partition is done, encryption of each record segment.is considered.* ***Results:*** *By then, sending of affirmation parts to various cloud servers is given most extreme significance. And when customer requires that information back, recovery of information from cloud servers and decoding of that information is gone ahead. In the wake of unscrambling, mix of that information and offer it to buyer is considered.* ***Conclusion:*** *The concept is to have the software promptly remit customer incorporation for customer's suppleness. The system which is capable to assure is the usage of half and half cryptography in a method which is more secure to send and accept the data.*

KEY WORDS
Cloud Computing, Reliability Algorithm, Subdivision Algorithm, Cryptography, Converging Methodology

Received: 15 May 2017
Accepted: 15 July 2017
Published: 20 Sept 2017

***Corresponding Author**
Email:
sprabubdu@gmail.com

INTRODUCTION

In the current season comprising of improvement, the Internet persuades the chance to be open late years, Distributed figuring is a web progression, utilized overall now-a-days to draw in the ultimate user to produce and employ programming lacking a complete consideration of accomplishment of the specific information from wherever at whatever point. Recollecting the genuine target which is to save monstrous information, server structures apply a few stockpiling substructures which are free from scale. The entire cloud space is shaped by combining these substructures together. The server space comprises of diverse reasons for engrossment which supports the usage of data. The cloud servers have parts of repetitive information [1]. The consumers can use unimportant evaluation of repository volume by maintaining a strategic distance from the copies. The server space has different parts to the clients like correspondence, archive space, numbers, exceptionally central data continues to get reflected, and so forth. Essentially, the data of client are secured in various stockpiling locales as neighboring servers including cloud [1].

The reliability of data is perceived as fundamental predicament within server space condition. Its adaptability for utilizing data wherever on the planet as the customers has the advances , approaches for getting to it has good fashioned security concern set away using a record can be balanced in various contraptions using an equivalent record. There are a touch of conflicting duplicates are open in scattered limit. [1]

The groupings of security concern are:

- Privacy, depicted confirmation that information is be kept puzzle.
- Trustworthiness, proposes for weakness to alter or beat information fortuitously or malpractice.
- Openness, which is the ability to get to that data at whatever point it is required subsequently there, is a requirement for a segment to deal with the security issues.
- A count that usages particularly secured transmission of information is to store and recoup from the cloud space are used. [2]

IBM Blue mix is the IBM open cloud manage that gives invaluable and web producers access to IBM programming for mix, security, exchange, and other key motivations behind control, and programming from business associates. Blue mix moderately has cloud plans that fit your needs. Despite whether you are a little business that approaches to scale, or an unending attempt that requires extra section, you can make in a cloud without edges, where you can relate your submitted relationship to people all around Blue mix affiliations open from IBM and their providers. All affiliations cases are supervised by IBM. Blue mix in like way gives middleware relationship to your applications to utilize. Blue mix gets up to speed for the application's motivation when it acquisitions new affiliations cases, and after that ties those relationship to the application. Your application can play out its certifiable occupation, leaving the relationship of the relationship to the establishment. As a rule, you don't need to push over the working structure and establishment layers when running applications on Blue mix. Layers, for example, root file systems and middleware areas are exceptional with the target that you can concentrate on your application code.

IIOAB JOURNAL
ISSN: 0976-3104

RELATED WORK

Complications of reliability in cloud computing

Interdependence & Corroboration: Cloud taking care of, subordinate upon the kind of cloud and besides the transmission appear, exhibited customers should immediately be secured and running with get to necessities and assents might be yielded in like way. This system is focus at scrutinizing and appreciative single cloud clients by utilize usernames and passwords assertions' to their cloud profiles Dispensation is a fundamental information reliability need in Cloud enrolling to ensure remission integrity progresses at the same rate. It takes after on in pertaining ascendancy and normal flexibilities over methodology streams inside Cloud figuring. Endorsement is kept up by cloud organization supplier.

Privacy in Cloud handling, security has certifiable effect particularly in supervising of cloud organization provider control over affiliations' information planned transversely finished different appropriated databases. It engages need when utilizing as a part of open in light of incontrovertible supremacy identity. Broadcasting protection of clients' outline and making their data reliable got the chance which considers the reliability of data customs to be affirmed [3].

For the two originalities utilizing cloud conditions and cloud suppliers, to make the data reliable, a dubious limitation is encoding. Vormetric encoding offers an unsophisticated methods for security containing key administration, fine-grained get to controls, and imaginative security knowledge information to keep delicate information very still inside various cloud situations i.e. open, private, mixture mists [4]. By means of encoding the cloud to use for cloud applications, one can meet understanding goals for encryption, flight of duties, and get to controls for secure data, including PCI-DSS and Data crosswise over Borders [5].

The dangers likewise incorporate those displayed by the disclosure of customer information to cloud suppliers and of information is conclusion as of the normal, blended information stockpiling used to sustenance cloud conditions. Additionally, distributed computing encryption(s) offers crude security insight on information access to encryption monitor data; such knowledge bolsters a Safety Data and Event Organization (SIEO) answer for recognizing dynamic diligent dangers or vindictive insiders [6].

Encryption provides a solitary, available elucidation that can just scramble any document, database, or accommodation wherever they live on kept up powerful and record frameworks, without prior application institution and keeping in mind that going around primary controlling trouble. Besides, cloud encryption contains nonstop key administration inside the clarification and is completely certain to applications and shoppers, consequently allowing existing strategies and strategy to persevere without any alterations [7].

The paper proficiently guards any information inside cloud situations. Correspondingly, the proposition of the cloud arrangement systems for upkeeps far reaching, strategy based separating of obligations is to present a perplexing level of reliability. The managers of cloud and the chiefs of origin and structure framework try to acquire illegal automatic ingress so that they can measure information. This can be counteracted and reasonable purchaser-implementation use can also be permitted.

Vision and Malthus, a secured information supplier of cloud, expels a danger that outstands personal information as well as information stockpiling that a solitary juncture has. They similarly proposed elucidations that point in high information in mists for simple joining with SIEO arrangements and to give finish information on utilize and get to. Get to endeavors utilizing SIEO arrangements enable computations to decide dangers to applications, and even executives. As far as encoding in cloud situations, Malthus enables associations that can screen if their data are shielded from the unhindered; it has diverse implementations and is a crossover cloud; conventional data file assets likewise exist on start [8].

A solitary, halfway expert course of action over all circumstances encourages the administration of cloud information security and in addition information security for physical and compelling datacenter properties. Notwithstanding, providers of cloud administrations make high esteem offices that have intensified information security offices in personal mists: SaaS, PaaS, IaaS, SaaS obliging, and a few others. Contemplating this, current encoding of cloud is a model arrangement as it is multi-inhabitant readied and versatile, is completed securely, and incorporates APIs and interfaces that are fundamental keeping in mind the end goal to work in collaboration with existing framework.

In June 2013, Edward Snowden disclosed the principle unpretentious components of the perception procedures of the NSA to the correspondents. With this, the field specialists foreseen that Snowden's exposure would unfairly impact cloud game plan masterminds. In August 2013, the DTIF said that the openings cloud would realize that the providers of cloud in the US lose 10% to 20% of the business to remote contenders overseas. In 2016, the DTIF said that the cloud providers would lose around 35 billion dollars in prospective arrangements. The CSA got a response from the European associations as to fears that the U.S. government is going to have induction to their information. In any case, following six months, the effect is all in all less extraordinary than had been typical. Regardless of reports of moderate offers of

ISSN: 0976-3104

the organizations of cloud by the US venders to abroad associations, and masters await the discharges uncovered by Edward Snowden are going to have small-scale impact on long-368 term bargains.

Encoding of Data: The breaks by Snowden pulled in a great deal of consideration inside the area of encoding. Thus, real administration contributors, as Microsoft, Yahoo, and Google, have since other coding to end-to-end information facilitating and administration for buyers. The present Google Cloud Storage mechanically encodes entirely pristine information recorded to plate. Server-side coding can be be offered in the blink of an eye for late information keep in Google mists.

Since the holes, Microsoft broadcasted its expectation to expand bolster for coding differed administrations like Outlook.com, Office 365, SkyDrive, and Windows Azure. By 2014, Microsoft hopes that can finish an occasion of standards for coding data exchanged betwixt purchaser locales and server farms and information in travel betwixt its individual information datum. Practically like Google, Microsoft wants to deal with information in shifted cloud specialist co-ops' mists.

Drop-down, Sonic.net, and Spider Oak have broadcasted bolster for comparative projects, encoding, alternatives, subsequent confirmation encoded information, and 2048-piece keys for the "ideal forward mystery" procedure. In accordance with specialists, these measures square measure essential for protecting the development of information between the shopper partnerships and thusly the providers of cloud administrations. Characterized reports from the NSA demonstrate that they're making an endeavor to debilitate coding calculations used by the overall population. The tap fiber interfaces that associate datacenters and repair providers offer the driving force for these endeavors.

Key administration and learning possession: in accordance with the United States, all through its debate with Lava bit, a protected email administrations provider, cloud benefit companies should fork over their coding keys once inquired. Such explanations have focused on sizeable consideration on key administration and learning possession. Eric Chiu, the leader of Trust, a cloud foundation administration organization, affirmed that however coding endeavors by benefit contributors caper a noteworthy half in up the reliability of cloud, their adequacy is prohibited.

Respectability: The validity essential care inside the server area, on an exceptionally fundamental extent in which the user gets to the data. In this way equality, stability, supervision are should make sure be viably compelled over all Cloud enlisting pass on models.

The orders utilized as a part of the Transportation Layer Security (TLS)/Protected Sockets Layer (SSL) strategy are represented. When TLS/SSL segments of anticipated structure had been nearness controlled, we test the division which is not used properly with complex salaries on chaperon as well as customer sides which occasional estimation has given as an outcome along with correspondence hold-ups which follow if complex as well as division tasks at hand change. At a point when the stated framework as well as calculation belonging are not used properly, the TLS/SSL yield is upgraded by abusing use of stated assets.

Exerting a chose pressure procedure to a TLS/SSL affiliation is not going to be ideal if the affiliation as well as calculation work region unit very surprising and dynamic. On the off chance that over the top information territory unit stacked for a meager data transmission TLS/SSL affiliation, pressure instrument intelligent setting heterogeneousness have to be connected. Notwithstanding, contemplating that normal TLS/SSL instruments give an unchanged pressure manner, realignment ask for by an implementation ought to change pressure algorithmic control that has to be connected. Consequently, component which licenses TLS/SSL to recognize the difficult pressure strategy for TLS/SSL associations in an exceedingly auspicious and clear way ought to be connected though considering the viewpoints given.

To begin with, we have a tendency to present firmly coupled, rib TLS/SSL composing, with the objective of boosting calculation and correspondence usage once TLS/SSL information sections territory unit sent and got. In run of the mill TLS/SSL, calculation schedules, similar to pressure and encoding operations likewise in light of the fact that the TLS/SSL organize schedules range unit dead in an exceedingly 369 approximately coupled way. This kind of implementation prompts visit impedance as well as stand performances inside TLS/SSL strategy.

Non-disavowal: Non-foreswearing in Cloud enrolling might be acquired by appertaining standard e-trade reliability customs as well as indicative workplaces to information transferal within the cloud acquisitions, for example, moved imprints, timestamps as well as proclamation proceeds associations.

The use of encoding and interpreting thoughts, to extend the protection of encrypted data that the customer transferred to the cloud server is examined in V.Masthanamma, G.Lakshmi Preya[9], The essential goal is to scramble and unscramble the information in a private way along weariness of lessened cash in encoding and unraveling process. Various amounts of keys will be made. So by repeating the strategy, keeping the strikes to extend the reliability of decoded data which the customer has transferred to the server in cloud has been made. The guideline intent is to scramble as well as decode the data in a shielded course with weariness in encoded as well as altering procedure. Various amounts of

keys are made as well as fundamental strikes are brought down. So by reiterating method it is utilized once more.

The paper of Mr. Akash Kanade et al [1] depicts that in the Partitioning Technique composing review is accomplished for data reputability examining, data stockpiling systems as well as encoding utensil. Dispensed arrangement is used so that the accessibility, data standards, uprightness of time tested accumulated organizations. The information limit using effectual data implementation method is utilized to accomplish diverse activities. Reliability examination is to encrypt the data by RSA.

The paper of Kiruthika R et al [10], describes figures , speak utilize distinctive pieces eradicates the keys in AES, which is a present drawback of DES and in AES. An execution relationship among celebrated calculations for different micro controllers shows that Advanced encryption standard has a PC cost of an indistinguishable demand from required for Triple Data encryption standard.

Another execution appraisal reveals that Advanced Encryption Standard has ideal position over figuring's 3Data encryption standard, Data encryption standard to the extent implementation period with different package magnitude as well as yield for encoding as well as furthermore unscrambling. Moreover by virtue of changing information sort, for instance, picture as opposed to alternate calculations.

The paper of Nasrin Khanezaei, Zurina Mohd Hanapi [2] portrays that various analyzes using encryption systems, distinctive strategies. Despite with the adjustment in the outcomes differentiating and the brisk improvement of distributed computing correspondences is taken care. Normally, reliability prototypes which are based on the cloud circumstances are isolated to confirmation prototypes, for instance, data affirmation prototypes, for instance, as well as get to organization prototypes, for instance. Using a blend of steganography encoding estimations, for instance, propelled encryption standard is one of the possible confirmation answers for securing circulated capacity organizations.

The fast development of the distributed computing market has likewise raised numerous genuine concerns with respect to information security and information administration, and these are thought to be the real boundaries to more extensive reception of cloud computing [11].

A study on security issues in benefit conveyance models of distributed computing, which concentrate more on characteristic reliability complications which have emanated due to the idea because of which the managing transference prototypes of a dispensed computing system [12.]

The paper [13] gives the investigation of information security issues and security insurance issues identified with distributed computing by keeping information access from unapproved clients, overseeing delicate information, giving exactness and consistency of information stored.

Sabarish et al., [14] have tended to various security challenges identified with cloud specialist co-op. Computerized stockpiling of PC information is pushing toward distributed computing which is an arrangement of framework gives information stockpiling to associations and people. Due to this huge scale, in the event that an assault happens in the system of a cloud it would be a major test to explore the cloud. [3]

The shifted security parts of security issues have been investigated thus proposes a structure to alleviate security issues at the sum verification and capacity level in distributed computing. Temperate security components should be sent by recommends that of coding, validation, and approval or by another philosophy to ensure the protection of customer's information on cloud storage [15].

PROPOSED METHOD

Encoding and decoding techniques

In order to make the encoded data of the customer reliable to the server in cloud is considered. The principal intent is to encode as well as translate data in a private utilizing short period as well as cash inencoding along with deciphering procedure. Various amounts of keys are to be be delivered including general assaults that are to be observed. This strategy encourages the user to let the assaults remain.

The strategies included are:

- Key creation
- Encoding
- Decoding

Encoding procedure

In the encoding procedure, a normal fonts in continuation of numbered modulo n. Using this, cipher text can be applied from plain text M. This is given like D=F g mod n,
Here D, is cipher text

F is word font
g is known key
D is unknown key

Decoding procedure

The backward procedure of encoding and decoding is decoding.
That is given as:
F= e mod n.
Where D =cipher text
F=word font
g =known key; D =unknown key

Partition Algorithm

Step 1: The input program along with its expanse is stacked.
Step 2: The program's expanse is examined.
Step 3: If expanse=invalid, state as invalid expanse.
Else expanse= s
Step 4: The program is divided into recorded and augmentation value.
Step 5: Send back files.

Merging algorithm

Step 1: Gather the fragments of decoded documents.
Step 2: Examine data progress.
Step 3: If data! Then not found
Else
Step 4: Margin quantity is taken
Step 5: Combine files.
Step 6: Back files.

Advanced Encryption Standard (AES):

The AES is a uniform key code in which the expanses of pieces are of 128 bits. The value is of can be of various bits.

Steps involved are:

- Ten rounds for hundred and twenty eight bits.
- Twelve rounds for hundred and ninety two bits.
- Fourteen rounds of two hundred and fifty six bits.

Rivest shamir adleman algorithm

It's an unsymmetric public key which is utilized to assist the encoding passwords. The encoding and decoding are carried out privately, hence G-mail, y- mail.

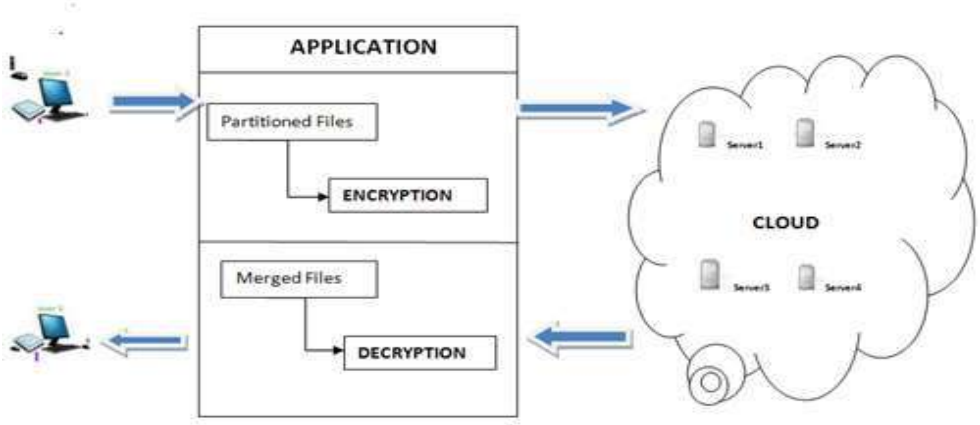


Fig. 1: Application architecture

The architecture of [Fig. 1] clarifies about how the engineering functions viably for the data to be exchanged in a protected way. In this context, the design incorporates customer, Server including software which is utilized to prepare the encoding as well as decoding procedures. In his context,

IIOAB JOURNAL
ISSN: 0976-3104

customer transfers the information's to the software in which the information is parceled as well as scrambled. At that point this information is transferred to the server in the cloud. This information is decoded, later combined, that are returned to the customer. It is finished along with proficient steganographic procedures in which utilization of uniform and lopsided calculations co-operatively that give greater reliability.

SYSTEM IMPLEMENTATION AND RESULTS

The plan of modules manages the accompanying:

Utilizer's Incorporation

In cloud innovation, the utilizer peruses the program by perusing interfaces. Client interfaces with this program to play out an errand. At the point, a program peruses a contribution from the client. After that, the program can interface with the environment in the cloud.

Partition

Client collaborates with program to parcel the information as well as save onto different servers. At the point at which the record went to the program that it is going to parcel. At the point at which a program peruses a document, it partitions as well as afterward connections distributed reserve. Later record gives demands as well as exchange to the coveted asset in the cloud.

Encoding

In this context, clients' encouraged to carry on the encoding procedure to give reliability divided records. The apportioned documents' are scrambled along with half breed steganography procedure use of uniform including hilter kilter calculations to give greater reliability.

Consolidation

The information is converged from servers which are later transferred to following procedure. The systems administration ideas in bluemix utilized here the combining procedure. Once the information is consolidated as well as finished along with decoding, the allotted period is going to be decreased.

Decoding

In this context, the clients' encouraged with the unscrambling procedure keeping in mind the end goal to give security to the combined records. The apportioned documents are decoded with the half and half cryptography system utilization of uniform as well as lopsided calculations which give greater reliability.

RESULTS

The application process includes creation of an application, edition of the code, source control and build and deployment. For this have to login to the IBM blue mix as given below:
a)

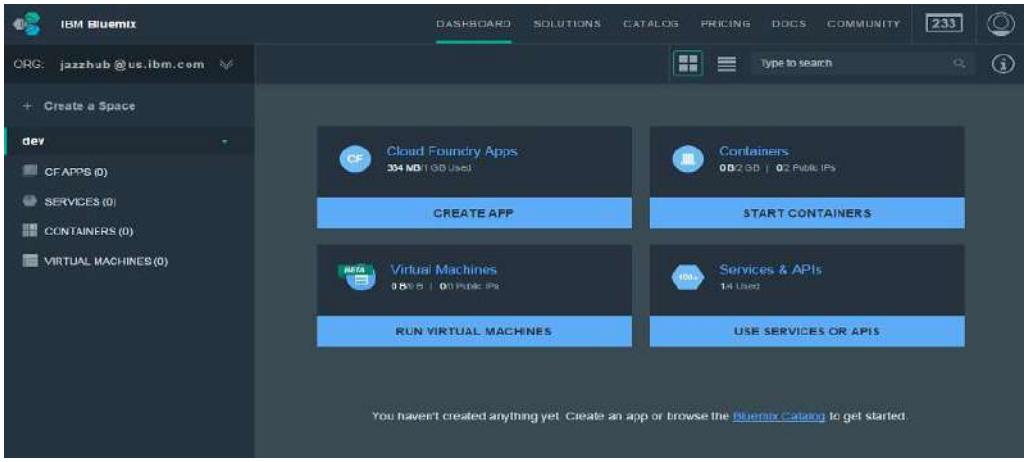


Fig. 2: Login to the blue mix page

The login procedure needs a pre-procedure of enrollment in the IBM blue mix. Through signing in the record, the dashboard in which the dashboard can be discovered that demonstrates the Cloud foundary applications, compartments, virtual machines and administrations.

b) Application can be created as shown below which deals with a couple of types of programs that could be created as well as picked when generating i.e., if for web or mobile application.

IIOAB JOURNAL
ISSN: 0976-3104

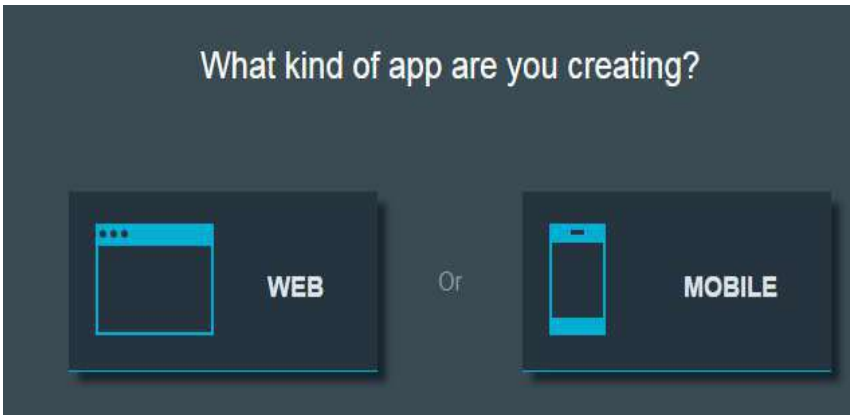


Fig. 3: Design of creating web or mobile application

c) Addition of GIT repository is made which is added and when EDIT code option is submitted to code the application with the functionalities that is required.

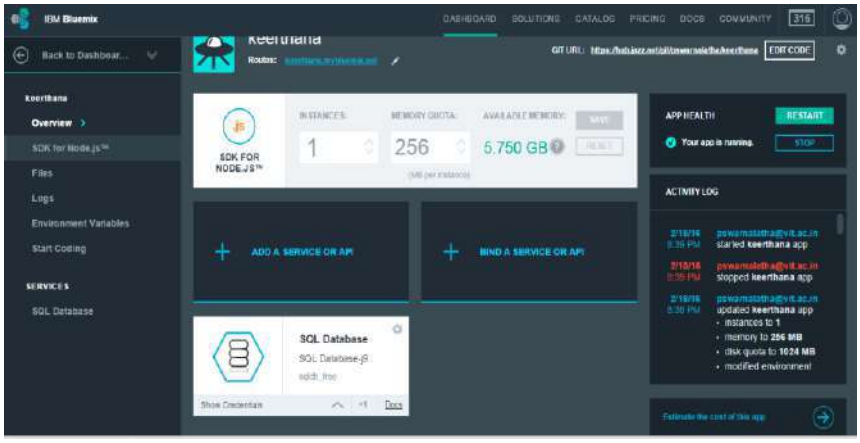


Fig. 4: Creation of an application with their operations created application with edit code.

CONCLUSION

A blend of shaky and adjusted encoding procedures (i.e. Rivert Shamir Adelman and Advanced Encryption typical systems) route for managing accomplish affirmation of cloud information separation. The significance was on Rivert Shamir Adleman encoding to offer inconvenience to aggressors and diminishment the period of information exchanging by using Advanced Encryption Standard encoding methodology. Methodology of transferring reports to server as well as recuperating documentations from cloud had been capable through uniform encoding independently. Using adjusted encoding technique, some portion will recoup the reports from the server which had been a consequence of the key dissemination problem. This problem gives a consummate out-turn in light of the way which is making different keys in a time span devouring. In like manner, the encryption technique winds up being twofold and dynamic, if there is extension of the report measure of two hundred and fifty six numbers. The measure of keys made for each record. The number finds the opportunity to be triple conditions for each estimation of reports set away in the cloud. As requirements could be an important problem that has to be managed for a liberal accumulated structure and moreover, the encoding as well as unraveling handle that done every two times each records structure overhead. A little while later, showed up contrastingly in connection to prevailing proficiency a hybrid approach for encoding, for example, it's more private to use. The specific shortcomings in future attempts will refresh the security of coursed taking care of associations.

CONFLICT OF INTEREST

There is no conflict of interest with any author regarding publication of this article

ACKNOWLEDGEMENTS

The authors of this paper express their gratitude to The Professor and Head, School of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirapalli Tamilnadu, India for providing guidance and support for this research work.

FINANCIAL DISCLOSURE

This research work was carried out without financial support from any organizations.

REFERENCES

[1] Kanade, Mr Akash, et al. [2015] Improving Cloud Security Using Data Partitioning And Encryption Technique, International Journal of Engineering Research and General Science. 3(1): ISSN 2091-2730.

[2] Khanezaei, Nasrin, Zurina Mohd Hanapi. [2014] A framework based on RSA and AES encryption algorithms for cloud computing services. Systems, Process and Control (ICSPC), 2014 IEEE Conference on. IEEE.

[3] Sulabha Patil, Uzma Ali, Dharaskar R V. [2015] Design and Development of System for detection of security Breach in Cloud environment, International Journal of Advance Research in Computer Science and Management Studies. 3(9):221-227.

[4] Modi C, Patel D, Borisaniya B, Patel A, Rajarajan M. [2013] A survey on security issues and solutions at different layers of Cloud computing, J. Super comput. 63(2):561–592.

[5] Jensen M, Schwenk J, Gruschka N, Iacono LL. [2009] On technical security issues in cloud computing. In 2009 IEEE International Conference on Cloud Computing. 109-116.

[6] Cloud Security Alliance. [2013] The Notorious Nine. Cloud Computing Top Threats in 2013, Security,1–14.

[7] Rodero-Merino L, Vaquero LM, Caron E, Muresan A, Desprez F. [2012] Building safe PaaS clouds: A survey on security in multitenant software platforms, Comput. Secur. 31(1):96–108.

[8] Jens-Matthias Bohli, Nils Gruschka, et al.[2013] Security and Privacy-Enhancing Multi cloud Architectures, IEEE Transactions on Dependable and Secure Computing. 10(4):212– 224.

[9] Masthanamma V, Lakshmi Preya G. [2015] An Efficient Data Security in Cloud Computing Using the RSA Encryption Process Algorithm, International Journal of Innovative Research in Science, Engineering and Technology.4:1441-1445.

[10] Pancholi, Vishal R, Bhadresh P Patel. [2016] Enhancement of Cloud Computing Security with Secure Data Storage using AES. International Journal for Innovative Research in Science and Technology. 2.9: 18-21

[11] Sharanjit Singh Er, Rasneet Kaur Chauhan Er. [2015] Introduction to CryptoCloud in Cloud Computing (IJETCAS) ISSN (Print): 2279-0047, ISSN (Online): 2279- 0055.

[12] Subashini, Kavitha V. [2011] A survey on security issues in service delivery models of cloud computing, Journal of Network and Computer Applications. 34:1–11.

[13] Jeevitha B K, Thriveni J, Venugopal K R. [2016] Data Storage Security and Privacy in Cloud Computing:A Comprehensive Survey,International Journal of Computer Applications (0975 – 8887).156(12).

[14] Srinivas J, Venkata K., Reddy S. and Qyser A. M. [2012] Cloud Computing Basics", International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 5, 2012,343-347.

[15] Deepika. [2017] Enhancement of Data Security for Cloud Environment Using Cryptography and Steganography Technique, International Journal of Innovative Research in Computer and Communication Engineering. 5(1):225-230.

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

REVIEW

BIG DATA IN HEALTHCARE – A REVIEW

Magesh G[1*], Swarnalatha P[2]

[1]*School of Information Technology and Engineering, VIT University, Vellore, INDIA*

[2]*School of Computer Science and Engineering, VIT University, Vellore, INDIA*

ABSTRACT

In exhibit world, information is developing step by step. Enormous information is created in a medicinal services association. Information is futile unless we won't locate any valuable data from it. Information examination turns into a test when we have such huge measure of information. In this paper, we have assessed recent year's papers. Information mining methodologies and planning for huge information is additionally surveyed. This paper will help future specialists to grow new methodologies and calculations to explain few difficulties in the healthcare services industry.

KEY WORDS
Big Data, Healthcare, Hadoop, Internet, Information Technology (IT)

INTRODUCTION

Finding talent is one of the most challenging tasks today, but is not usually prevailing because of the problem, not being technical. We are living in a world where everyone looks for profits. According to most of the industries, most of the students graduating from India are not fit for work and thus they go through about a year of industrial training, shouldn't all colleges be responsible for that? I think all colleges should take care of that field training, instead of relying on industries. Most students can learn programming on their own and if not, should be specially taken care of, but for the rest, logic and algorithms should be the focus of their study, applying and integrating it in field should be the primary objective.

The next in line is the expensiveness of the server hardware, for the time being, we cannot reduce the costs immediately, but we can make a smart investment, and may be reduce the costs in the long run. I believe we should work on real time compression techniques, thus improving its efficiency and file size. Taking this into account, an intermediate system should be integrated with the original system between the server and desired application, here map reduce, dedicated for file compression and decompression. During the first stages of implementation we might not notice a huge difference, but in the long run a significant reduction in server load may be noticed.

Received: 11 May 2017
Accepted: 10 July 2017
Published: 21 Sept 2017

There are very few studies regarding the wearing and laundering of lab coats in hospitals and medical practice. This study highlights the role of lab coats acting as vector for transmitting health care infections to the patients and the common areas where contamination occurs.

Literature Review

Analyzing patients sentiments using hadoop and machine learning

Prescient demonstrating is basic to changing substantial clinical informational indexes, or "enormous clinical information," into significant learning for different medicinal services applications. Machine learning is a noteworthy prescient demonstrating approach; however, two hindrances make its utilization in social insurance testing. Initial, a machine learning instrument client must pick a calculation and dole out at least one model parameters called hyper-parameters before demonstrate preparing. The calculation and hyper-parameter esteem utilized regularly affect demonstrate precision by more than 40 %, yet their determination requires many works concentrated manual cycles that can be troublesome notwithstanding for PC researchers. Second, numerous clinical qualities are more than once recorded after some time, requiring worldly accumulation before prescient demonstrating can be performed. Much work serious manual emphasis are required to distinguish a decent combine of total period and administrator for each clinical characteristic. The two obstructions result in time and human asset bottlenecks and block medicinal services overseers and specialists from soliciting an arrangement from imagining a scenario in which questions while testing chances to utilize prescient models to enhance results and lessen costs. This paper portrays their plan of and vision for PredicT-ML (forecast instrument utilizing machine taking in), a product framework that means to defeat these boundaries and mechanize machine learning model working with enormous clinical information [Fig-2]. The paper introduces the point by point plan of PredicT-ML. Foresee ML will open the utilization of enormous clinical information to a huge number of social insurance managers and analysts and increment the capacity to progress clinical research and enhance human services. [1]

The possibility that the simply phenomenological information that they can separate by breaking down a lot of information can be valuable in medicinal services appears to negate the craving of VPH analysts to construct itemized unthinking models for singular patients. Yet, practically speaking no model is ever altogether phenomenological or totally unthinking. They propose in this position paper that enormous information investigation can be effectively joined with VPH advancements to create hearty and compelling

***Corresponding Author**
Email: magesh.g11@gmail.com

in silico prescription arrangements. So as to do this, huge information innovations must be additionally created to adapt to some particular pre requisites that rise up out of this application. Such necessities are: working with delicate information; investigation of mind boggling and heterogeneous information spaces, including non-printed data; circulated information administration under security and execution limitations; specific examination to incorporate bioinformatics and frameworks science data with clinical perceptions at tissue, organ and living beings scales; and concentrated investigation to characterize the "physiological envelope" amid the everyday life of every patient. These area particular prerequisites propose a requirement for focused financing, in which huge information advancements for in silico medication turns into the exploration need. [2]

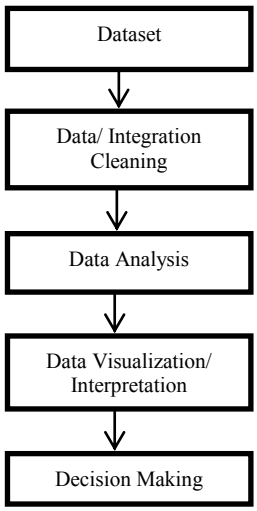


Fig. 1: Decision-making

With the advancement of savvy gadgets and distributed computing, increasingly general wellbeing information can be gathered from different sources and can be examined in an extraordinary way. The immense social and scholarly effect of such improvements caused an overall buzz for huge information. In this audit article, they compressed the most recent uses of Big Data in wellbeing sciences, incorporating the suggestion frameworks in social insurance, web based pestilence observation, sensor-based wellbeing conditions and sustenance security checking, Genome-Wide Association Studies (GWAS) and articulation Quantitative Trait Loci (eQTL), surmising air quality utilizing huge information and metabolomics and bionomics for nutritionists. They additionally checked on the most recent advancements of enormous information gathering, stockpiling, exchanging, and the cutting edge logical strategies, for example, Hadoop circulated document framework, MapReduce, proposal framework, profound learning and system Analysis. Finally, they talked about the future points of view of wellbeing sciences in the time of Big Data. [3]

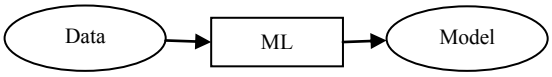


Fig. 2: Machine learning

This target of this paper to depict the guarantee and capability of enormous information examination in human services. The paper portrays the beginning field of enormous information examination in human services, talks about the advantages, diagrams a structural system and strategy, depicts illustrations revealed in the writing, quickly examines the difficulties, and offers conclusions. Results: The paper gives a wide diagram of huge information investigation for medicinal services analysts and experts. Conclusion is that enormous information examination in human services is developing into a promising field for giving understanding from huge informational indexes and enhancing results while diminishing expenses. Its potential is extraordinary; however there remain difficulties to overcome. As large information investigation turns out to be more standard, issues, for example, ensuring protection, defending security, setting up principles and administration, and constantly enhancing the devices and advancements will gather consideration. Enormous information examination and applications in medicinal services are at an incipient phase of improvement, however quick advances in stages and instruments can quicken their developing procedure. [4]

This work shows a novel, Cloud-based framework for information gathering and capacity to catch smoking conduct with e-cigarette. A client's smoking information created by the day by day utilization of e-cigarettes

IIOAB JOURNAL
ISSN: 0976-3104

is transferred to the cloud through versatile web and a Bluetooth association between an advanced mobile phone and the e-cigarette. All individual character can be scrambled and an investigation personality number will be doled out to each subject for information security insurance. The remote stage in the cloud can give productive investigative execution on an enormous volume of information with high speed of information creation. Information mining on smoking conduct will better comprehend the methods for utilizing the e-cigarette. This information framework will likewise be conceivably utilized as a part of other epidemiological investigations in general wellbeing. In this investigation, they display their work in the improvement of an information catching stage in the cloud for e-cigarettes associated with cell phones. This stage is intended for future examination on the utilization of e-cigarette for smoking lessening or restraint. Information on smoking conduct and the viability of e-cigarette utilize is caught by a cell phone APP, transmitted to the cloud, and a synopsis provide details regarding smoking status will be naturally created for get to through the cell phone. They trust that this intelligent stage on e-cigarette utilize will build the achievement rate of smoking restraint. [5]

In this paper creators portray the principal ever endeavor of constant information investigation for social insurance datasets utilizing as a part of memory databases and they benchmark and analyze two such in-memory database frameworks to examine responsiveness and capacity to deal with multifaceted nature of run of the mill wellbeing information investigation undertakings, they share their work in advance outcomes and blueprint key issues that should be tended to for approaching advances in this imperative enormous information vertical. This paper investigates persuading situations for investigating exceptionally complex human services informational collections continuously by performing specially appointed inquiries. They look into two economically accessible In-memory database advances Mem SQL and Volt DB for responsive medicinal services information investigation. Both of these databases are New SQL social databases. For correlation they utilized Medicare Claims Synthetic information from Center for Medicare and Medicaid Services (CMS). The primary commitments of this work are speedier information investigation and empowering ongoing expectations. [6]

An expanding number of the elderly populace wish to carry on with an autonomous way of life, as opposed to depend on meddling consideration programs. A major information arrangement is displayed utilizing wearable sensors fit for completing nonstop observing of the elderly, cautioning the significant guardians when vital and sending relevant data to a major information framework for examination. A test for such an answer is the improvement of setting mindfulness through the multidimensional, dynamic and nonlinear sensor readings that have a powerless relationship with detectable human practices and wellbeing conditions. To address this test, a wearable sensor framework with an astute information forwarder is talked about in this paper. Ping Jiang et al. displayed an arrangement of utilizing wearable sensors fit for ceaseless observing of the elderly and sending the information to the huge information frameworks. A major information framework oversees high volume, speed and assortment of data from various sources, to address this test a wearable sensor framework with keen information forwarder was presented receiving Hidden Markov display for human conduct acknowledgment for nourishing important information to the framework. The sensor readings and conditions of a client are sent to Big Data investigation for enhancing and customizing the nature of care. [7]

This paper concentrates on enormous information issues, points of view and arrangements as for human services industry. Diverse enormous information arrangements and a structure for taking care of heterogeneous clinical information has been examined which if actualized successfully could demonstrate supportive in basic leadership for understanding consideration and general medicinal services results The social insurance industry is good to go for change with a specific end goal to enhance quiet care and to make it savvy by utilizing on most recent advances. Digitization has prompted huge measure of advanced information particularly in human services industry. This paper proposes a technique for taking care of heterogeneous human services information utilizing the correct innovation and design which can possibly change social insurance results and the nature of patient care at ideal cost. This paper likewise shows different logical instruments that can be utilized to use profits by the immense arrangement of social insurance information. Legitimate choice of devices to do examination on medicinal services information can give promising outcomes. The distinctive reception challenges highlighted in the paper are a portion of the open issues and can be considered as a territory of research for future work. [8]

In this paper, creators propose a Genetic Algorithm based scheduler for such Big Data Cloud where decoupled computational and information administrations are offered as administrations. The approach depends on developmental techniques concentrated on information conditions, computational assets and powerful use of data transfer capacity in this way accomplishing higher throughputs. They portray a booking model in view of GA and the model is assessed and contrasted and before works, for example, coordinate making and different heuristics systems through the reproduced information. They build up a planning system, where computational assets and information stockpiles are decoupled with the information recreated over capacity storehouses which are geological scattered. Here, the issue is centered on gathering the occupations in view of the information necessities, and the goal is to limit the aggregate make traverse considering both computational assets and correspondence transmission

capacity viably. The proposed family/gather booking model tends to the information concentrated issues to limit the turnaround time of the employments where the registering and information assets are decoupled. [9]

Predictive big data analytics

This paper talks about the part of Big Data in the present medicinal services segments. The change of utilizing complex advancements by medicinal services suppliers to pick up bits of knowledge from clinical datasets and settle on educated choices had changed by Big Data Analytics. With the assistance of Hadoop, the objective of successful human services administration can be accomplished by giving compelling information driven administrations to individuals by anticipating their requirements This paper gives an outline of putting away and recovery strategies, Big Data apparatuses and methods utilized as a part of social insurance mists, part of Big Data Analytics in medicinal services and talks about the advantages, viewpoints in early fields of prescient investigation, confronts challenges and gives arrangements. The method for association of information after extraction from various layers and coordination of it is additionally a testing undertaking. Overseeing and keeping up the human services information is smidgen basic since social insurance information is kept up in medicinal services records, they are colossal in nature and need much space, further, on the off chance that they require related information it needs to experience with examination to get significant information. The part of technocrats in social insurance field ought to keep up adjust amongst business and innovation, which clear a way for Big Data in medicinal services data frameworks. Applying Big Data in medicinal services space made an upheaval in keeping up and managing intricate, unstructured heterogeneous information. [10]

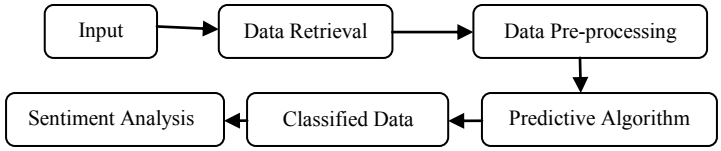


Fig. 3: Sentiment analysis

This paper portrays the outline, usage of fluffy master framework for conclusion of diabetes. The parts of fluffy master framework are fuzzification interface, Fuzzy evaluation technique and De-fuzzification interface. Fuzzification interface changes over the fresh esteems into fluffy esteems. Fluffy evaluation strategy utilizes fluffy administrators, enrollment work, connection fluffy rationale and likelihood to oversee uncertainty in rules. De-fuzzification interface changes over the subsequent fluffy set into fresh esteems. To show the viability of the proposed calculation MATLAB Fuzzy Logic tool kit is utilized for execution evaluation. The outcome shows that the fluffy evaluation system is exceptionally compelling in enhancing the exactness for diabetes application. [11]

Master frameworks are late result of manmade brainpower. It is an arrangement of projects that control encoded information to take care of issues in a particular area. Diabetes is a constant ailment that requires persistent medicinal care in tolerant self-administration training to avoid intense difficulties. Gestational diabetes mellitus (GDM) is a kind of Diabetes display in around 3-4% of all pregnancies. This paper proposes a technique to distinguish the GDM in pregnant ladies. Analysts contributed their work in diagnosing the diabetes mellitus. Their proposed work concentrates on displaying a specialist framework to analyze GDM utilizing Feed forward Neural Network design. [12]

This paper manages utilizing fluffy rationale to limit vulnerability impacts in reconnaissance. It thinks about the origination of a productive fluffy master framework that had two qualities: bland and strong to vulnerabilities. Dissecting separation between factors ideal and genuine esteems is the primary thought of the examination. Fluffy induction framework chooses, at that point, about noteworthy factors state: ordinary or irregular. A correlation between three proposed fluffy master frameworks is displayed to highlight the impact of participation number and sort. Next to, being bland this framework could likewise be connected in three fields: modern observation, camera reconnaissance and medicinal observation. To uncover brings about these fields, MATLAB is utilized to understand this approach and to reproduce frameworks reactions which uncovered intrigued conclusions. [13]

Choice emotionally supportive networks give quiet particular suggestions to mind suppliers amid clinical experiences. The MobiGuide extend additionally addresses patients by settling on them clients of the choice emotionally supportive network through their Smartphone interface, by customizing direction as per the patients' inclinations in a way touchy to their own specific circumstance, and by including patients in a common basic leadership together with their care suppliers. Furthermore, the MobiGuide framework gives choice help in non-clinically-controlled situations, for example, the patient's working environment and home, and can convey some of its proposals even without a care-supplier's intercession; consequently, it is a Ubiquitous Guidance System. The clinical learning that constitutes the rationale of the universal direction

framework depends on confirm based clinical rules. These properties make the MobiGuide framework possibly exceedingly available and quick. Also, patients' wellbeing is expanded. While keeping most checked patients out of the facility, the framework prompts an expansion in wellbeing quality and a lessening in medicinal services costs. [14]

Patient-centric clinical decision support system

Clinical practice rules (CPGs) expect to enhance the nature of care, decrease unjustified practice varieties and lessen social insurance costs. With the end goal for them to be viable, clinical rules should be coordinated with the care stream and give persistent particular exhortation when and where required. Henceforth, their formalization as PC interpretable rules (CIGs) makes it conceivable to create CIG-based choice emotionally supportive networks (DSSs), which have a superior shot of affecting clinician conduct than story rules. This paper audits the writing on CIG-related systems since the initiation of CIGs, while centering and drawing subjects for characterizing CIG explore from CIG-related productions in the Journal of Biomedical Informatics (JBI). The subjects traverse the whole life-cycle of CIG advancement and include: learning obtaining and determination for enhanced CIG configuration, including (1) CIG demonstrating dialects and (2) CIG procurement and particular techniques, (3) mix of CIGs with electronic wellbeing records (EHRs) and authoritative work process, (4) CIG approval and check, (5) CIG execution motors and strong devices, (6) exemption taking care of in CIGs, (7) CIG support, including investigating clinician's consistence to CIG suggestions and CIG forming and development, lastly (8) CIG sharing. I analyze the transient patterns in CIG-related research and examine extra subjects that were not distinguished in JBI papers, including existing topics, for example, defeating execution boundaries, demonstrating clinical objectives, and worldly articulations, and in addition advanced topics, for example, tolerant driven CIGs and circulated CIGs. [15]

Enthusiasm for distinguishing individuals at danger of creating Type 2 Diabetes Mellitus (T2DM) has increased extraordinary significance particularly at its asymptomatic stage, when early intercessions have a demonstrated helpful impact on clinically important results. The current indicative criteria are centered on recognizing bunches with fundamentally expanded commonness of small scale vascular difficulties. This proposes the current indicative strategies are feeling the loss of the chance to recognize IGT and IFG (pre-diabetes) and early side effects of T2DM, which prompts a late ID and treatment of patients and the ensuing improvement of inconveniences, which could be stayed away from with a prior mediation. This paper presents the MOSAIC venture which means to enhance the present principles for diabetes conclusion and administration. [16]

Diabetes is an incessant sickness that requires a man with diabetes to make a huge number of day by day self-administration choices and perform complex care exercises. Diabetes self-administration training and support (DSME/S) gives the establishment to help individuals with diabetes to explore these choices and exercises and has been appeared to enhance wellbeing outcomes.1-7 Diabetes self-administration instruction (DSME) is the way toward encouraging the information, expertise, and capacity essential for diabetes self-mind. Diabetes self-administration bolster (DSMS) alludes to the help that is required for executing and maintaining adapting abilities and practices expected to self-oversee on a continuous premise (see advance definitions in Table 1). Albeit diverse individuals from the social insurance group and group can add to this procedure, it is vital for medicinal services suppliers and their training settings to have the assets and an efficient referral procedure to guarantee that patients with sort 2 diabetes get both DSME and DSMS in a steady way. The underlying DSME is regularly given by a wellbeing proficient, while continuous help can be given by work force inside training and an assortment of group based assets. DSME/S programs are intended to address the patient's wellbeing convictions, social needs, current learning, physical restrictions, enthusiastic concerns, family bolster, budgetary status, medicinal history, wellbeing education, numeracy, and different elements that impact every individual's capacity to address the difficulties of self-administration. [17]

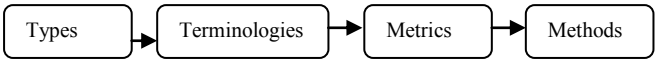


Fig. 4: Clinical decision support system.

Meeting the perplexing needs of patients with perpetual disease or weakness is the single most noteworthy test confronting sorted out therapeutic practice. Common care is not doing the occupation; many reviews and reviews have uncovered that sizable extents of incessantly sick patients are not accepting viable treatment, have poor malady control, and are miserable with their care (1). Consequences of randomized trials likewise demonstrate that compelling malady administration projects can accomplish significantly preferred results over normal care, the control mediation. These trials, alongside the thoughts and endeavors for development talked about in this issue, demonstrate that they can enhance care and results. As the articles recommend, these upgrades won't come effectively. In the event that they are to enhance administer to most patients with interminable disease, the proof firmly recommends that they

IIOAB JOURNAL
ISSN: 0976-3104

reshape their walking look after this reason. Essential care hone was to a great extent intended to give prepared get to and care to patients with intense, shifted issues, with an accentuation on triage and patient stream; short arrangements; analysis and treatment of manifestations and signs; dependence on research center examinations and solutions; brief, instructive patient training; and patient started development. Patients and families battling with interminable ailment have diverse necessities, and these requirements are probably not going to be met by an intense care association and culture. They require arranged, customary collaborations with their parental figures, with an emphasis on capacity and counteractive action of intensifications and complexities. This connection incorporates precise appraisals, thoughtfulness regarding treatment rules, and behaviorally advanced help for the patient's part as self-chief. These associations must be connected through time by clinically significant data frameworks and proceeding with follow-up started by the medicinal practice. [18]

Numerous wellbeing frameworks in Canada are starting to recognize the need to enhance nurture people with endless sickness. Their present framework is basically intended for intense care; dire indications regularly trump the need or longing to bring the patient's incessant infection under ideal administration. As the Institute of Medicine's (IOM) turning point report "Intersection the Quality Chasm" recommended, investing more energy or accomplishing business as usual won't work and just genuine framework change may endeavor to conquer the apparently inconceivable inadequacies in the system.1 because of this suggestion to take action, the Chronic Care Model (CCM) was produced as a guide to enhance the nature of interminable ailment mind. The CCM does not recommend a handy solution, but rather a multi-dimensional, efficient way to deal with a mind boggling issue. [19]

Decision making in healthcare

Basic leadership in doctor's facilities has developed from being assessment construct to being situated in light of sound logical proof. This basic leadership is perceived as proof based practice. Ceaseless production of new confirmation consolidated with the requests of consistently hone makes it troublesome for wellbeing experts to stay up with the latest. Clinical pathways are archive based devices that give a connection between the best accessible proof and clinical practice. They give suggestions, procedures and time allotments for the administration of particular medicinal conditions or intercessions. Clinical pathways have been actualized worldwide however the confirmation about their effect from single trials is conflicting. This survey expected to abridge the confirmation and evaluate the impact of clinical pathways on proficient practice (e.g. nature of documentation), understanding results (e.g. mortality, complexities), and length of healing center stay and doctor's facility costs. Twenty-seven examinations including 11,398 members were incorporated for investigation. The fundamental outcomes were lessening in-healing facility intricacies and enhanced documentation related with clinical pathways. Entanglements evaluated included injury diseases, draining and pneumonia. Most investigations revealed a diminished length of stay and decrease in healing facility costs when clinical pathways were executed. Significant variety in ponder outline and settings counteracted factual pooling of results for length of stay and healing center expenses. By and large poor revealing kept the recognizable proof of qualities regular to fruitful clinical pathways. The creators reasoned that clinical pathways are related with diminished in-healing facility intricacies. [20]

CONCLUSION

Enormous information in health insurance is the rising examination field; this paper is attempting to introduce a diagram of huge information investigation, its applications in different fields. This paper likewise draws out the difficulties and the exploration openings in huge information.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Gang Luo. PredicT ML: [2016] a tool for automating machine learning model building with big clinical data. Luo Health Inf Sci Syst 4:5.

[2] Marco Viceconti, Peter Hunter, Rod Hose. [2015] Big data, big knowledge: big data for personalized healthcare. IEEE Journal of Biomedical and Health Informatics .

[3] Tao Huang, Liang Lan, Xuexian Fang, Peng An, Junxia Min, Fudi Wang. [2015] Promises and Challenges of Big Data Computing in Health Sciences. Big Data Research.

[4] Wullianallur Raghupathi, Viju Raghupathi. [2014] Big data analytics in healthcare: promise and potential. Health Information Science and Systems.

[5] Kelvin KF Tsoi, Yong-Hong Kuo, Helen M. Meng. [2015] A Data Capturing Platform in the Cloud for Behavioral Analysis Among Smokers. IEEE International Congress on Big Data.

[6] Muaz Mian, Ankur Teredesai, David Hazel, Sreenivasulu Pokuri, Krishna Uppala. [2014] Work in Progress - In-

Memory Analysis for Healthcare Big Data. IEEE International Congress on Big Data.
[7] Ping Jiang, Jonathan Winkley, Can Zhao, Robert Munnoch, Geyong Min, Laurence T Yang. [2014] An Intelligent Information Forwarder for Healthcare Big Data Systems with Distributed Wearable Sensors. IEEE SYSTEMS JOURNAL.
[8] Prabha Susy Mathew, Anitha S Pillai. [2015] Big Data Solutions in Healthcare: Problems and Perspectives. ICIIECS.
[9] Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, Rajkumar Buyya. [2014] Genetic Algorithm based Data-aware Group Scheduling for Big Data Clouds. IEEE/ACM International Symposium on Big Data Computing.
[10] Rishika Reddy A, Suresh Kumar P. [2016] Predictive Big Data Analytics in Healthcare. Second International Conference on Computational Intelligence & Communication Technology.
[11] Aronson A R. [2001] Effective mapping of biomedical text to the UMLS Metathesaurus: the Meta Map program. Proceedings / AMIA Annual Symposium. AMIA Symposium, pages 17–21.
[12] Ayers SL, Kronenfeld JJ. [2007] Chronic illness and health-seeking information on the Internet. Health (London), 11(3):327–347.
[13] Baeza-Yates R, Calder´on-Benavides L, Gonz´alez-Caro C. [2006] The intention behind web queries. In Proceedings of the 13th International Conference on String Processing and Information Retrieval, SPIRE'06, Berlin, Heidelberg. Springer-Verlag. 98–109.
[14] Bhavnani SK, Jacob RT, Nardine J, Peck FA. [2003] Exploring the distribution of online healthcare information. In CHI '03 Extended Abstracts on Human Factors in Computing Systems, CHI EA '03, New York, NY, USA, ACM. 816–817.
[15] Boot CR, Meijman FJ. [2010] Classifying health questions asked by the public using the icpc-2 classification and ataxonomy of generic clinical questions: an empirical exploration of the feasibility. Health communication. 25(2):175–181.
[16] Boser BE, Guyon IM, Vapnik VN. [1992] A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory. ACM. 144–152
[17] Broder A. [2002] A taxonomy of web search. SIGIR Forum, 36(2):3–10.
[18] Cartright M A, White R W, Horvitz E. [2011] Intentions and attention in exploratory health search. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, New York, NY, USA. ACM. 65–74.
[19] Chee B, Berlin R, Schatz B. [2009] Measuring population health using personal health messages. In AMIA Annual Symposium Proceedings. American Medical Informatics Association.92.
[20] Chen L, Zhang D, Levene M. [2013] Question retrieval with user intent. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, New York, NY, USA.ACM.973-976.
[21] Chen L, Zhang D, Mark L. [2012] Understanding user intent in community question answering. In Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, New York, NY, USA. ACM.823-828.
[22] Chih-Chung Chang and Chih-Jen Lin. LIBSVM – A Library for Support
[23] Vector Machines, April 2013.
[24] Cho J H, Liao V Q, Jiang Y, Schatz B R. [2013] Aggregating personal health messages for scalable comparative effectiveness research. In Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, BCB'13, New York, NY, USA. ACM.907:907–907:916.
[25] Cohen J. [1960] A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 20(1):37.
[26] De Choudhury M, Morris M R, White R W. [2014] Seeking and sharing health information online: Comparing search engines and social media. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems, ACM.1365–1376.

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

AN IMPROVED MECHANISM FOR USER PROFILING AND RECOMMENDATION USING CASE-BASED REASONING

Bhavithra Janakiraman[1*], Saradha Arumugam[2], Aiswarya Jayaprakash[1]

[1]*Department of Computer Science and Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, INDIA*

[2]*Department of Computer Science and Engineering, Institute of Road and Transport Technology, Erode, Tamilnadu, INDIA*

ABSTRACT

Background: *Web Page recommendation system is the process of identifying suitable webpages that matches with user query and recommending them for further access. Today most of the research works have been focusing on applying machine learning techniques for improving the accuracy of recommendation process. Such algorithms predict suitable webpages by analyzing the past search experience of the currently active user. These past search history will be stored in the form of user profiles.* ***Methods:*** *This paper proposes an improved mechanism for developing user profiles using selective usage-based attributes. Attributes such as Page Rank, Page Weight, Bounce Rate, Exit Rate and Conversion Rate of each webpage visited by the corresponding user will be computed and stored in the profile of corresponding user. Case-Based Reasoning is further applied for clustering the user profiles based on similar search interest and generation of cluster summary. The subsequent webpages suitable for the currently active user is predicted based on cluster summary.* ***Results:*** *Experiments were conducted on eight categories of datasets comprising of benchmark and real-time web log data. The performance was analyzed in terms of accuracy and mean absolute error rate.* ***Conclusions:*** *Results expresses that, the proposed method of user profiling and clustering outperforms existing algorithms with improved accuracy and lower error rate.*

KEY WORDS
User Profile, Case-Based Reasoning, Bounce Rate, Exit Rate, Conversion Rate

Received: 12 May 2017
Accepted: 12 July 2017
Published: 21 Sept 2017

***Corresponding Author**
Email: bavi.rr@gmail.com
Tel.: 9443170753

INTRODUCTION

Current information era tries to engage plenty of online users by providing appropriate search results based on user query. Unfortunately, such results do not satisfy all the users of various categories. Information is to be retrieved from the web based on individual user needs. In other words, query search results must be customized accordingly to provide user satisfaction. Recommendation system is the thriving research area today, in which personalization is done to analyze user's search interest and provide better results even for those users who do not reveal their search interest explicitly [1]. Collaborative filtering approach provides recommendations based on other users who have similar interest and preferences. To support personalization in recommendation [2, 3, 4], it is necessary to record user's search history [5,6]. Web users may provide their search interest input either explicitly or implicitly [7]. Explicit input includes showing their interest by providing user-rating or subscription for articles and messages in the corresponding web sites. Implicit input are those that were given indirectly. Hence intelligent algorithms are to be executed to achieve those implicit inputs.

Web log mining is basically divided into three major categories [8] namely, Web Usage Mining, Web Structure Mining and Web Content Mining. Web Usage Mining focus on analyzing the URLs (web pages) visited by a normal user. In terms of analyzing the usage of a web page, each user's personal interest of that web page and overall search intention can be predicted. This type of mining plays a vital role in predicting the implicit interest of a user in web page recommendation system. Web structure mining is the process of analyzing the hyperlink structure within a web page. Web content mining is another predominant area of research where the actual search topic of the user is analyzed. Content based similarity should be considered with more weightage for any effective recommendation system.

This research mainly targets on analysis of web usage log and the content log of frequent web surfers. User profile is constructed based on URLs visited by users from various IP addresses. This paper proposes an improved mechanism to generate user profile using such implicit inputs. The User profile will be created for individual users using attributes with respect to web usages such as Page Rank (PR), Page Weight (PW), Bounce Rate (BR), Exit Rate (ER) and Conversion Rate (CR). Also with these attributes, web content based attribute such as the frequency of keywords present in each of these corresponding pages will also be recorded in the user profile.

The following are some of the contributions of the proposed paper:

- An improved mechanism has been suggested for profiling user's search interest using Usage-Based attributes such as PR, PW, BR, ER and CR along with Content-Based attributes such as most frequent keywords.
- Clustering these user profiles initially using k-means and followed incrementally through Case-Based Reasoning (CBR). Incremental clustering approach is followed in order to dynamically cluster the users based on varying search interest and pattern. CBR is adapted suitably to generate a summary profile for each cluster [9]. This summary comprises of the mean values of Usage-based and content-based attributes.
- Finally, the recommendation system obtains the user query and recommends the web pages only based on matching cluster summary. Thus optimization is achieved with good accuracy and better response time.

IIOAB JOURNAL
ISSN: 0976-3104

The rest of this paper is organized as follows. In section II, related work to this paper has been discussed. Section III discusses about the User profiling mechanism. In Section IV, Case Based Reasoning is applied to effectively cluster the users based on profile attributes which produce cluster summary. Section V discusses about how web pages are recommended through matching cluster summaries. Section VI covers results and discussion. In section VII, final conclusion and future work is discussed

RELATED WORK

Predictions of web pages that could likely be visited by end users are recommended based on the user interest and previous navigation history. Various traditional methods such as collaborative filtering, association rules, clustering, sequential patterns, hybrid methods and semantic web [10, 11] are used for such predictions. Collaborative filtering [1] is one of the most common approaches used for providing recommendation by finding similar users. Pearson correlation coefficient and cosine based approach can be used to find similar users [12]. This traditional approach can still be improved by applying normal recovery Collaborative filtering [13]. But recommendation done using pure collaborative filtering approach may lead to problems such as popularity bias, cold start problem, handling dynamic pages etc. Case Based Reasoning generates the user profile and uses similarity knowledge to predict relevant profiles for the currently active user [14]. Such profile includes Page Rank [15] as a major feature which is computed using HITS and Page Rank algorithm [10].

Collaborative Filtering

Collaborative filtering is one of the most common approaches used for recommendation. Collaborative Filtering systems collect visitor opinions on a set of objects using ratings, explicitly provided by the users or implicitly computed. In explicit ratings, users assign ratings to items or web pages, or a positive (or negative) vote to some web pages or documents [10]. The implicit ratings are computed by considering the access to a Web page. A rating matrix is constructed where each row represents a user and each column represents an item or web page keywords [13]. Items could be any type of online information resources or products in an online community such as web pages, videos, music tracks, photos, academic papers, books etc. Collaborative filtering systems predict a particular user's interest in an item using the rating matrix. Alternatively, the item-item matrix, which contains the pair-wise similarities of items, can be used as the rating matrix. Rating matrix is the basis of CF methods. The ratings collected by the system may be of both implicit and explicit forms. Although CF techniques based on implicit rating are available for recommendation, most of the CF approaches are developed for recommending items where users can provide their preferences with explicit ratings to items

Content-based filtering

Content-based filtering is a type of information extraction system, where web pages are extracted based on the semantic similarity between the content in those web pages visited by users in past history [16, 17]. Web content mining applications mostly rely on content-based filtering approaches. Content-based filtering offers predominant support for web page recommendation system. In this technique, the keywords and its frequency of occurrence in those web pages that were previously visited are collected. Then, the semantic similarity between such keywords will be analyzed for further process. For example, consider two users "u1" and "u2" who frequently visit web pages based on their domain of interest [18]. Let u1 always focus on heath related web pages and u2 focus on gadget-related sites. Now, during the real time if any active academic user "ua" search for the query "apple", he will be mostly related to apple devices based sites, rather than apple fruit. So, he will be recommended the sites referred by u2. Similarly, when a dietician "ub" searched for "apple" he will be recommended the sites referred by u1. Recommendation engine classifies "ua" as an academic user and "ub" as a dietician based on the contents (keywords) of the web pages navigated in past history.

Case-based reasoning

Case Based Reasoning [14, 19, 20] is a process of finding solutions to new problems based on the solutions of similar past problems. This approach of refining new solutions can be considered as a clustering problem, where based on previous clustering structure and defined attributes, any new case can generally be grouped to any one of the previous structure. As the analysis of the new case is totally based on the previous cases and experiences, this approach generally provides good accuracy and cluster purity. Today many researchers work in the area of applying case-based reasoning in web mining concepts.

USER PROFILE GENERATION

This paper proposes an improved mechanism for gathering user's implicit search interest based on his previous search pattern and history [2, 21]. The following attributes are used for recording the implicit interest of any user by analyzing URLs visited by the corresponding user:

- Page Rank (PR)
- Page Weight (PW)
- Bounce Rate (BR)
- Exit Rate (ER)

- Conversion Rate (CR)
- Keywords (KEY)

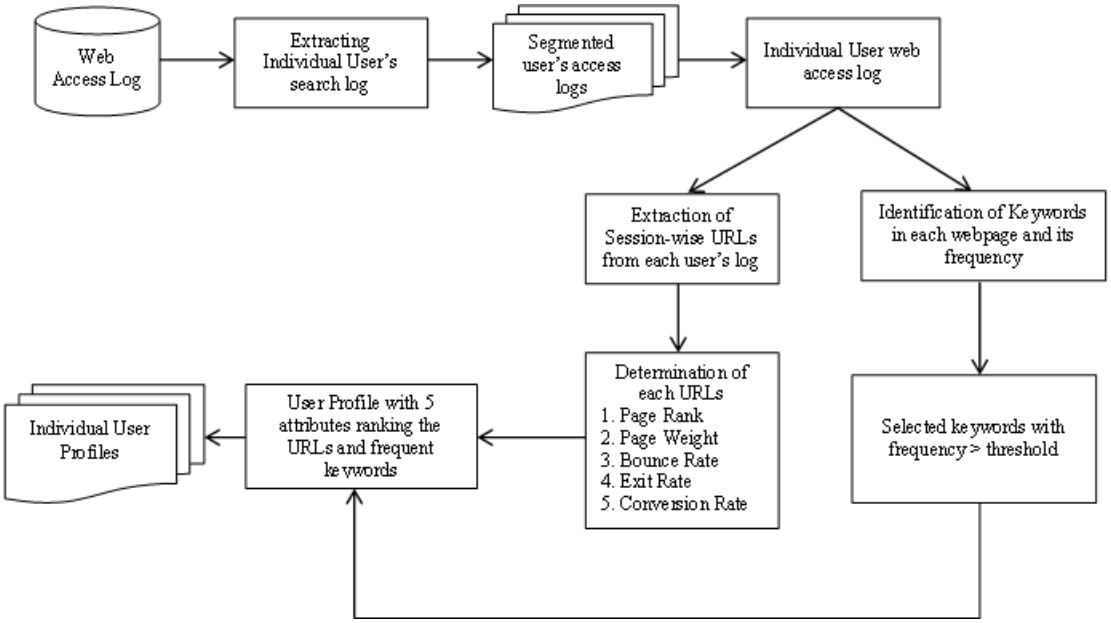


Fig. 1: Overall architecture of user profile generation from web access log dataset.

The overall architecture of User-Profile generation system is shown in [Fig. 1]. The following paragraph discusses about the procedures used for user profile generation.

Page rank (PR)

Page Rank is a numerical value that measure's a webpage importance among the group of similar web pages [22]. Such page rank is computed based on Random Surfer model [23]. This algorithm computes the page rank based on the link structure of the web page [23]. A page gets hold of high rank if the addition of the ranks of its backlinks is high. The rank of the given page is thus computed using the following equation (1)

$$PR = \frac{1}{N}\left[(1-d) + d\sum_{v \in B(u)} Page_{Wt} \times \frac{PR(v)}{N_v}\right] \quad (1)$$

Where, u represents a web page. B(u) is the set of pages that point to u. PR(v) is the page rank of page v that points to page u. Nv is the number of outgoing links of page and d is the damping factor that is set between 0 and 1. The damping factor is the decay factor that represents the chance of a user stop clicking links within a current page and then requesting another random page [22]. $Page_{Wt}$ is termed as Page weight which is calculated based on frequency and duration as in equation (2).

Page weight (PW)

Page weight is calculated based on frequency and duration as in Equation (2).

$$Page_{Wt}(PW) = NV(pi) \times \tau pi \quad (2)$$

Where τpi is total time spent by the user on particular web page. A quick jump might also occur due to the short length of a web page so the size of page may affect the actual visiting time. Hence, duration is normalized by the length of the web page, i.e. the total bytes of the page. NV(pi) is the number of times that a page is accessed by different users; computed by equation (3).

$$NV(pi) = \sum_{j=1}^{N}\sum_{pi \in sj} n \quad (3)$$

Where,

$$n = \begin{cases} 0, & \text{if pi is not present atleast once in Session Sj} \\ 1, & \text{if pi is present atleast once in Session Sj} \end{cases}$$

Bounce rate (BR)

The web page access percentage with respect to session wise grouping of access pattern is called as bounce rate. Today BR plays a vital role in web analytics. Web pages access pattern is grouped into sessions based on date and time difference between two consecutive page requests. If the date and time difference is exceeding the certain time limit of 10 minutes, we have grouped the access patterns as clusters called as sessions. The page pi's access rate between all such sessions is computed as BR using the equation (4)

$$BR(pi) = \frac{\sum_{s \in S} \sum_{(pi \in s) \cap (i=1)} \tau pi}{NS(pi)} \times TS \quad (4)$$

Where 's' represents each session from the complete set of sessions 'S' and TS represents the total number of sessions active by a web user. NS(pi) denotes the total number of sessions where page pi has been accessed.

Exit rate (ER)

The rate at which, the web page (pi) will be at the end of the session is computed as ER. Here, the occurrence of pi being the last entry within the session is calculated to identify the exit rate using the equation (5)

$$ER(pi) = \frac{\sum_{s \in S} \sum_{(pi \in s) \cap (i=N)} \tau pi}{NS(pi)} \times TS \quad (5)$$

Conversion Rate (CR)

The conversion rate for each web page is computed as the ratio between total sessions accessed by a user to the total number of sessions that contains the page pi. Equation (6) computes the conversion rate of page pi.

$$CR(pi) = \frac{TS}{NS(pi)} \times 100 \quad (6)$$

Keywords (KEY)

The set of keywords present in all the visited web pages is collected together as a local dictionary for that corresponding user. The set of identical words after performing tokenization, removal of stop words and stemming is generally called as Keywords. A matrix as shown in [Table 1] is developed after keywords identification. The matrix is populated with term frequency which corresponds to the number of times a keyword (term) occurs within a particular web page (denoted by its URL). Pruning operation is also performed through the elimination of those keywords that are below the certain threshold level. The threshold level is computed using equation (7). Finally the top "n" keywords from "k" web pages are selected for further clustering process.

$$Thershold = \frac{1}{k} \sum_{i=1}^{k} Term\_Freq(KEYi) \quad (7)$$

Table 1: Keyword matrix comprising of frequency of keywords in local dictionary of top k webpages

URLs	KEY1	KEY2	KEY3	KEY4	…	KEYN
URL 1	12	5	13	16	…	7
URL 2	4	17	16	7		14
URL 3	17	14	12	2		1
URL 4	11	15	8	18		12
.....						
URL K	10	13	5	1	…	12

CLUSTERING USING CASE-BASED REASONING

Case-based reasoning is then applied for effective clustering of user profiles based on the similarity in their search and access pattern. For example, users those who are always searching for product based web pages will be clustered together. Similarly, users those who search for "Web Mining" based research articles will be united into a single cluster. Such similarities among the search histories were identified using keyword match. The following algorithm (1) applies k-means method of clustering for initially segmenting the users. The initial set of clustering will result in a matrix as shown in [Table 2]. In order to incrementally cluster/ re-cluster the active members, case-based reasoning is used. The final cluster summary generated using CBR approach is shown in [Table 3].

Algorithm 1: Applying Case-Based Reasoning for clustering user profiles

Input: User profiles with Usage-based features such as PR,PW,BR,ER,CR and Keywords(1..N)
Output: Cluster summary comprising of keywords(1..N) and Top URLs for further recommendation
Initialize: N← Max_KEY; K←No_of_UserProfiles; M← No_of_Clusters; c[1..M]← No_of_users in cluster

Begin
1. Retrieve all user profiles and store as Hash Table
   1.1. UP_KEY[i] ← [K1i,K2i,K3i, … ,KNi]
   1.2. UP_URL[i] ← [PRi,PWi,BRi,ERi,CRi]
2. Initially cluster user profiles based on keywords using k-means algorithm
   2.1. Randomly assign 'K' users into 'M' clusters; Such that each cluster contains 'c' users (where c ← K/M)
   2.2. For all clusters c = 1 to M, identify the mean of term_frequency for each keyword KEY = 1 to N
   2.3. Compare individual user profile's KEY[1..N] term_frequency with KEY_MEAN[1..N] of that corresponding cluster
   2.4. Move the user profile which has the nearest match between its KEY[1..N] term_frequency with KEY_MEAN[1..N]
   2.5. Repeat from step 2.2 until there are no more moves
   2.6. Return initial level of clusters.
3. Apply Case_Based Reasoning for further clustering which includes updated/new user profiles
   3.1. Identify the mean value for all usage-based features resulting in Avg(PR), Avg(PW), Avg(BR), Avg(ER), Avg(CR)
   3.2. Compare the Keyword similarity of updated/new user profile with cluster's mean.
   3.3. Identify top 'T' clusters that nearly matches with updated/new user profile
   3.4. Compare the updated/new user profile's usage-features with those 'T' clusters
   3.5. Move the updated/new user profile to the corresponding cluster with the best match
4. Identification of Cluster Summary
   4.1. For each cluster c from 1 to M repeat the following
   4.2. Sort the cluster members based on usage-based features such as PR,PW,BR,ER,CR
   4.3. List the top 'S' User profile's Keywords (1..N) and their URL's as the corresponding cluster's summary.
   4.4. Return the summary of all clusters c[1..M].
End

Table 2: User profile matrix used for clustering using case-based reasoning (an example)

User_ID	PR	PW	BR	ER	CR	KEY1	…	KEYN	Cluster_ID
User 1	3	32	58	45	4.67	12		9	3
User 2	2	21	41	38	2.45	44		23	1
User 3	5	14	68	23	7.54	17		6	2
User 4	17	12	32	18	3.56	11		3	3
User 5	1	37	67	34	9.87	16		13	2
User 6	8	45	58	24	1.34	11	…	13	3
User 7	1	32	32	12	3.67	34		15	1
User 8	4	11	18	37	5.24	45		32	1
User 9	2	21	44	43	6.13	23		12	2
User 10	3	37	24	23	8.44	22		9	4
……									
User K	16	22	54	32	3.59	19	…	7	4

Table 3: Cluster summary generated using CBR algorithm (cluster comprising of book searchers profiles)

KEY1	…	KEYN	Top URLs visited by users in the corresponding cluster
44		23	www.googreads.com,www.amazon.com, www.infibeam.com
34		15	www.openiningthebook.com
17		13	www.quora.com, www.bookadventure.com
16	…	13	www.educatorstechnology.com
12		9	www.readingrockets.org
12		6	www.magicbox.com
11		2	www.kidsreads.com

RECOMMENDATION PROCESS

This is the only online activity done by web servers as soon as end user provides search query. The cluster summary generated using CBR approach will be used for further recommendation. As the user provides the query through online, information retrieval and further recommendation must be obtained with quick response time. Hence recommendation algorithms should concentrate both on accuracy and as well as good response time. The recommendation process of the proposed system uses cluster summary in order

to efficiently predict the next web page that will be possibility expected by the currently active user. The overall recommendation process is given in [Fig. 2].

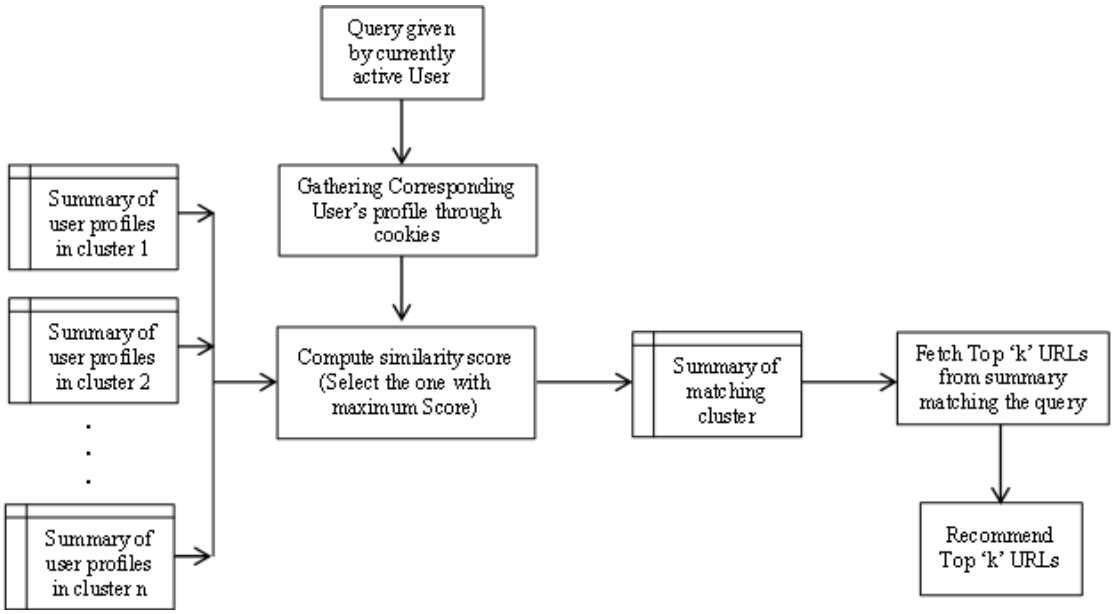


Fig. 2: Recommendation of web pages based on similarity between active user and cluster summary

Recommendation system fetches the profile of the corresponding user who provided the search query. These user profiles can be either saved within web servers, search engines or even at the client side itself. If the profile is stored within the client end, cookies are used for fetching the profiles. One of the issues faced here would be that if end user deletes the cookie files, recommendation accuracy may be reduced. After successfully retrieving the profile, the cluster summary of the corresponding cluster where the user is intended to be grouped is identified. From the list of URLs stored in the summary, top 'k' URLs will be recommended for the currently active user. If the recommended web pages are selected by the end users, the page's PR, PW and BR will automatically raise favoring the corresponding web page to be given with high weightage to be considered for the next time. Also, if the user does not visit a page, its view, weight and bounce rate will automatically diminish again favoring the recommendation process by lowering the selection probability of low-rated web pages to be recommended.

RESULTS

Experiments were conducted using eight categories of datasets including benchmark web server repository called AOL web access log dataset and real time web access log from Dr.Mahalingam College of Engineering and Technology (MCET), Pollachi, Tamilnadu, India. The web access log dataset is divided into eight samples of equal size with 50 records as mentioned in [Table 4]. The descriptions of these datasets have been discussed below:

AOL Web Access Log (AOL)

AOL web access log dataset [24] was used. The log file contains web query log data from ~650k users. In order to have privacy preservation, IP addresses of individual users were anonymized. Hence each user is represented by unique ID. The schema of this log dataset is: {AnonID, Query, Query Time, Item Rank, ClickURL}. Where, AnonID denotes an anonymous user ID number; Query denotes to the query issued by the user; Query Time refers the time at which the query was submitted for search; Item Rank states that if the user clicked on a search result, the rank of the item on which they clicked is listed; Click URL states that if the user clicked on a search result, the domain portion of the URL in the clicked result is listed.

MCET Web Access Log (MCET)

The log file was collected from students browsing history, given by Dr.Mahalingam College of Engineering & Technology, Pollachi, Tamilnadu, India. The size of the dataset used was about 23,709 KB with 1 lakh entries, consists of 77 different IP addresses, date & time of visiting the web pages, method URL/protocol, status, received byte and connection type. Each webpage may include advertisements, pictures, videos, textual content etc. The banned and invalid URLs are ignored during web content extraction. The preprocessed log will contain IP Address, date & time and URL.

Table 4: Various sample datasets used for experimentation process

Data Set Title	Description
AOL_SET 1	The top 50 users who access web frequently were selected based on the maximum length (no. of URLs) within each session.
AOL_SET 2	The top 50 users who do not access web frequently were selected based on the minimum length (no. of URLs) within each session.
AOL_SET 3	The top 50 users having profile with maximum number of identical search keywords were selected.
AOL_SET 4	The top 50 users having profile with minimum number of identical search keywords were selected.
MCET_SET 1	Without any conditions, access log of 50 users were selected randomly
MCET_SET 2	Uniform sampling was performed to select one user after each 50 records.
MCET_SET 3	The query was analysed and categorized into various domains. 50 users accessed under academic category were selected.
MCET_SET 4	The query was analysed and categorized into various domains. 50 users accessed under entertainment category were selected.

The model is trained using both AOL dataset and real time MCET dataset covering the samples excluding the eight sets mentioned in [Table 4]. Testing experiments were conducted with traditional Collaborative Filtering (CF) algorithm and the proposed User Profiling and Recommendation using CBR (UPR-CBR) approach using these eight samples.

Evaluation metrics

The performance of existing (CF) and proposed (UPR-CBR) systems are evaluated using Precision, Recall, F-Measure [16] and Mean Absolute Error (MAE) [13]. Let 'R' denote the total number of web pages in the collection, 'A' denotes set of web pages that are answered and 'Ra' denotes Set of relevant web pages that are retrieved. Precision is calculated as shown in equation (8) which is defined as the ratio of a number of relevant web pages retrieved to the total number of answered web pages. Recall is defined as the ratio of a number of relevant web pages retrieved to the total number of web pages. It is expressed as a percentage and calculated as shown in Equation (9). F-Measure is the harmonic mean of precision and recall, the F-measure or balanced F-score is calculated as shown in equation (10). Mean Absolute error (MAE) is the average absolute deviations of predictions to the ground truth values. It measures the deviation of actual value and predicted value using the equation (11)

$$\text{precision} = \frac{|R_a|}{|A|} \tag{8}$$

$$\text{Recall} = \frac{|R_a|}{|R|} \tag{9}$$

$$\text{Harmonic Mean} = \frac{2}{\frac{1}{r} + \frac{1}{p}} \tag{10}$$

$$\text{MAE} = \frac{\sum_{u,i} |r_{u,i} - \hat{r}_{u,i}|}{N} \tag{11}$$

Where,

- r denotes recall
- p denotes precision
- $r_{u,i}$ denotes actual accuracy of webpage i observed by the user u
- $\hat{r}_{u,i}$ denotes the predicted value of webpage I for user u
- N denotes the total number of predicted value

The following [Table 5] compares the performance of Collaborative Filtering (CF) algorithm and User-Profiling based Recommendation using Case-Based Reasoning (UPR-CBR) approaches in terms of accuracy and error rate as defined in equations (11) and (12) respectively. It has been found that in all the samples of datasets both in benchmark access logs and real time MCET access logs, the proposed system outperforms existing collaborative recommendation. The error rate is also found to be low for UPR-CBR in all datasets. Multiple experiments were conducted with varying cluster values (M). An average F1-measure of all experiments under the same sample set is summarized in [Table 5].

Table 5: Performance analysis of algorithms in terms of accuracy and error rate

Training Dataset	Accuracy (Harmonic Mean)		Mean Absolute Error (MAE)	
	UPR-CBR	CF	UPR-CBR	CF
AOL_SET1	70	56	0.41	0.54
AOL_SET2	66	61	0.47	0.52
AOL_SET3	68	57	0.47	0.58
AOL_SET4	68	56	0.46	0.58
MCET_SET1	64	60	0.48	0.53
MCET_SET2	71	63	0.46	0.55
MCET_SET3	71	66	0.43	0.47
MCET_SET4	76	73	0.36	0.40

CONCLUSION

This paper proposes an improved mechanism of profiling user's web pages of interest through their search history and navigation by using usage-based features such as Page Rank, Page Weight, Bounce Rate, Exit Rate and Conversion Rate. The frequency of keywords in the web pages visited by corresponding users was recorded to identify the similarity among user community for effective recommendation. Hence proposed system's user profiling mechanism includes both collaborative approach and content-based filtering systems. Case-based reasoning was then applied for clustering the user profiles among vast set of user community with similar web navigation history. CBR concludes the clustering by providing a cluster summary comprising of top level web pages with high usage-based features and content-based features. During online information retrieval as soon as the user provides the search query, the corresponding user profile is just compared with the matching cluster summary and web pages are recommended with quick response time. The usage-based and content-based attributes will be dynamically updated favoring relevant pages to be recommended with high probability and the probability of considering irrelevant pages will be reduced gradually based on user's click rate. Experiments were conducted using eight categories of datasets from benchmark and real time log entries. Results infer that the performance of proposed system is found to be improved in terms of accuracy and error rate.

CONFLICT OF INTEREST

There is no conflict of interest with any author regarding publication of this article

ACKNOWLEDGEMENTS

The authors of this paper express their gratitude to the respected management team and network administration team of Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India for providing the web access log files of the college for this research in order to perform real time testing.

FINANCIAL DISCLOSURE

This research work was carried out without financial support from any organizations.

REFERENCES

[1] Zibin Zheng, Hao Ma, Michael Lyu R, Irwin King. [2009] Wsrec: A Collaborative Filtering Based Web Service Recommender System, IEEE International Conference on Web Services, 437-444.

[2] Sarabjot Singh Anand, Bamshad Mobasher. [2005] Intelligent Techniques for Web Personalization, Springer-Verlag Berlin Heidelberg, 1-36.

[3] Abdul Manan Ahmad, Mohd. Hanafi Ahmad Hijazi. [2004], Web Page Recommendation Model for Web Personalization, Springer-Verlag Berlin Heidelberg, 587–593.

[4] Magdalini Eirinaki, Charalampos Lampos, Stratos Paulakis, Michalis Vazirgiannis. [2004] Web Personalization Integrating Content Semantics and Navigational Patterns, WIDM'04, ACM.

[5] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, [2013] A New Algorithm for Inferring User Search Goals with Feedback Sessions, IEEE Transactions on Knowledge and Data Engineering, 25(3): 502-513.

[6] Dimitrios Pierrakos, Georgios Paliouras. [2010] Personalizing Web Directories with the AIDOF Web Usage Data, IEEE Transactions on Knowledge and Data Engineering, 22(9): 1331-1344.

[7] Şule Gunduz-Oguducu, Tamer Ozsu, M. [2006] Incremental Click-Stream Tree Model: Learning from New Users for Web Page Prediction, Distributed Parallel Databases, Springer Science, 19: 5–27.

[8] Şule Gunduz - Oguducu. [2010] Web Page Recommendation Models: Theory and Algorithms, Synthesis Lectures on Data Management, 2: 1-85.

[9] Yahya Al Murtadha, Md. Nasir Bin Sulaiman, Norwati Mustapha, Nurlzura Udzir. [2010] Improved Web Page Recommendation System using Profile Aggregation Based on Clustering of Transactions, American Journal of Applied Sciences, 8(3):277-283.

[10] Pooja Devi, Ashlesha Gupta, Ashutosh Dixit. [2014] Comparative Study of HITS and PageRank Link Based Ranking Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, 3: 5749-5754.

[11] Murat Goksedef, Şule Gunduz-Oguducu. [2010], Combination of Web Page Recommender Systems, Elsevier Journal on Expert Systems with Applications, 2911–2922.

[12] Zibin Zheng, Hao Ma, Michael Lyu, R, Irwin Kin. [2011] QoS-aware Web Service Recommendation by Collaborative

Filtering, IEEE Transactions on Services Computing, 4:140-152
[13] Huifeng Sun, Zibin Zheng, Junliang Chen, Michael Lyu R. [2012] Personalized Web Service Recommendation via Normal Recovery Collaborative Filtering, IEEE Transactions on Services Computing, 6: 573-579.
[14] Yong-Bin Kang, Shonali Krishnaswamy, Arkady Zaslavsky. [2014], A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge, IEEE Transactions on Cybernetics, 44: 473- 487.
[15] Huifeng Sun, Yong Peng, Junliang Chen, Chuanchang Liu, Yuzhuo Sun. [2011] A New Similarity Measure Based on Adjusted Euclidean Distance for Memory-Based Collaborative Filtering, Journal of Software, 6:993-1000.
[16] Niki, R, Kapadia, Kinjal Patel. [2012] Web Content Mining Techniques – A Comprehensive Survey, International Journal of Research in Engineering & Applied Sciences, 2(2)1869, 1877.
[17] Lina Yao, Quan Z Sheng, Anne HH Ngu, Jian Yu, Aviv Segev. [2015] Unified Collaborative and Content-Based Web Service Recommendation, IEEE Transactions on Services Computing, 8(3): 453-466.
[18] Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu. [2014] Web-Page Recommendation Based on Web Usage and Domain Knowledge, IEEE Transactions on Knowledge and Data Engineering, 26(10): 2574-2587.
[19] Daxa K. Patel. [2014], A Retrieval Strategy for Case-Based Reasoning using USIMSCAR for Hierarchical Case, International Journal of Advanced Engineering Research and Technology, 2(2): 65-69.
[20] Shweta Tyagi, Kamal, K Bharadwaj. [2012] A Hybrid Recommender System Using Rule-Based and Case-Based Reasoning, International Journal of Information and Electronics Engineering, 2(4): 586-590.
[21] Aditi Shrivastava, Nitin Shukla. [2012] Extracting Knowledge from User Access Logs, International Journal of Scientific and Research Publications, 2(4)
[22] Pooja Devi, Ashlesha Gupta and Ashutosh Dixit [2014] Comparative Study of HITS and PageRank Link Based Ranking Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, 3: 5749-5754.
[23] Ciobanu D, Dinuca CE. [2012] Predicting the Next Page that will be visited by a Web Surfer using Page Rank Algorithm, International Journal of Computers and Communications, 6(1): 60-67.
[24] Michael G Noll. [2006] AOL Research Publishes 650,000 User Queries, 2006, http://www.michael-noll.com/blog/2006/08/07/aol-research-publishes-500k-user-queries/
[25] Yahya Al Murtadha, MD Nasir Bin Sulaiman, Yahya Norwati Mustapha, Nur Izura Udzir, Zaiton Muda. [2011] Web Page Recommendation System for Anonymous Users Based on Web Usage Mining, Advances in Communications, Computers, Systems, Circuits and Devices, ISBN: 978-960-474-250-9.
[26] Tranos Zuva, Sunday, O. Ojo, Seleman M. Ngwira, Keneilwe Zuva. [2012] A Survey of Recommender Systems Techniques: Challenges and Evaluation Metrics, International Journal of Emerging Technology and Advanced Engineering, 2(11): 382-386.
[27] Wesley Chu, Tsau Young Lin. [2005] Foundations and Advances in Data Mining - Studies in Fuzziness and Soft Computing, Springer Verlag, vol. 180.
[28] Pearl Pu, Li Chen, Rong Hu. [2012] Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art, Springer Science 22: 317–355.
[29] Schröder G, Thiele M, Lehner W. [2011] Setting Goals and Choosing Metrics for Recommender System Evaluations, Proceedings of the Second Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI 2).

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

ENHANCED PERFORMANCE IN WORK FLOW SCHEDULING USING GTTD

V. Balaji*, P. Swarnalatha

Department of SCOPE, VIT University, Vellore, Tamilnadu, INDIA

ABSTRACT

Researchers in each research team share their information and procedure distributed resources for directing their experimentations. These tests are actuality complemented in association with groups that are internationally distributed. Information area difficult and transmitted information overhead are main challenges for development such Information –demanding technical workflow application in cloud computing. These requests are important to the period of big data and task execution involves incontrollable and manufacturing huge amount of input/output information with data dependencies amongst responsibilities. Grouping Technique based Task Dependency | (GTTD) to decrease performance overhead and to increase the computational granularity of technical workflow tasks is obtainable in this paper. And this paper suggests the Information-intensive workflow development system to reduce make distance of the Information-intensive workflow applications, which can be exhibited as a focussed acyclic graph. Grouping technique is authenticated by using imitation based exploration though Work flow Sim.

KEY WORDS
application software,scheduling, grouping, cloud

Received: 29 May 2017
Accepted: 27 July 2017
Published: 22 Sept 2017

***Corresponding Author**
Email:
vuppala.balaji@gmail.com
Tel.: +91-8688778087

INTRODUCTION

A workflow could be a high-level specification of a collection of tasks that represent procedure science or business flows and therefore the dependencies between the tasks that has got to be glad so as to accomplish a particular goal. The business progress is sometimes management flow-driven and includes constructs to specify methods, conditions, and should embrace human interaction. Typically, a business workflow implements a company's product or service. Scientific workflows usually touch upon massive an outsized quantity of knowledge and advanced calculations and so utilize large storage capacities and computing resources. During a scientific progress, a task is associate degree workable program with a collection of input parameters and files. Scientific workflows are generally information flow-driven and don't have made management flow structures, whereas notable exceptions exist, like Askalon [1]

In several analysis fields, particularly in lay to rest disciplinary field like bioinformatics and climate simulation, scientific workflows are typically each computation and data-intensive. Workflow running typically wants large-scale computing resources however additionally large storage. Today, scientific workflows applications carries with it many thousands of tasks, consume gigabytes or terabytes input information sets and generate similar amounts of intermediate data. These applications are noted as data-intensive progress applications. A knowledge intensive application consists of applications that manufacture, manipulate, or analyze information within the vary of many megabytes (MB) to petabytes (PB) [2]. Having the power [3] to makeover workflows between execution resources needs a precise degree of flexibility once it involves information placement and information movement. Most workflows need a knowledge storage resource near the execution web site so as to execute in associate degree economical manner. The benefits of cost-effectiveness, on-demand resource provision and straight forward for sharing.

A workflow could be a high-level specification of a collection of tasks that represent procedure science or business flows and therefore the dependencies between the tasks that has got to be glad so as to accomplish a particular goal. The business progress is sometimes management flow-driven and includes constructs to specify methods, conditions, and should embrace human interaction. Typically, a business workflow implements a company's product or service. Scientific workflows usually touch upon massive an outsized quantity of knowledge and advanced calculations and so utilize large storage capacities and computing resources. During a scientific progress, a task is associate degree workable program with a collection of input parameters and files. Scientific workflows are generally information flow-driven and don't have made management flow structures, whereas notable exceptions exist, like Askalon [1].

In several analysis fields, particularly in lay to rest disciplinary field like bioinformatics and climate simulation, scientific workflows are typically each computation and data-intensive. Workflow running typically wants large-scale computing resources however additionally large storage. Today, scientific workflows applications carries with it many thousands of tasks, consume gigabytes or terabytes input information sets and generate similar amounts of intermediate data. These applications are noted as data-intensive progress applications. a knowledge intensive application consists of applications that manufacture, manipulate, or analyze information within the vary of many megabytes (MB) to petabytes (PB) [2]. Having the power [3] to makeover workflows between execution resources needs a precise degree of flexibility once it involves information placement and information movement. Most workflows need a knowledge storage resource near the execution web site so as to execute in associate degree economical manner. The benefits of cost-effectiveness, on-demand resource provision and straightforward for sharing Distributed computing has developed in prominence with research group for conveying work flows. At the point when work flows are broadly performed in cloud situations that comprise of various data centres, there is a critical requirement for creating systems which can put the application information

IIOAB JOURNAL
ISSN: 0976-3104

crosswise over all-inclusive circulated data centres and plan assignments as per the information design to lessen both the dormancy and make span for work flow execution. Make span is a term alluding to the execution metric of work flow and characterized as the interim between the begin time of first errand and the end time of definite assignment [4].

Regularly, an information serious application work flow is booked to limit add up to information exchange time as well as cost, storage room utilized, add up to execution time or potentially a blend of these. In such conditions, information area and exchanged information overhead are essential difficulties for information escalated applications to decrease execution time and size of exchanged information. Overlooking area of information supports high data transmission utilization cost [5]. For a few applications [6] up to 90 % of the execution time is spent on document exchange. Numerous scientists [2] have proposed a few components for exchanging information with the goal that information exchange time is limited. These strategies are: information parallelism, information gushing, and information throttling. Information throttling is a procedure of depicting and controlling when and at what information rate is to be moved rather than moving information starting with one area then onto the next as right on time as could be allowed. Planning and execution overhead are high when low execution of fine-grained assignments is a typical issue in broadly disseminated stages. This paper proposes errand grouping strategy to limit these overhead in light of assignment reliance.

Information development and exchange overhead are not enormous issues in a little group condition. Be that as it may, logical work flow framework is intended for researchers to participate over a few server farms. In runtime of execution, information development and exchanged time can effect on aggregate execution of work flow application. "Moving information to a server farm will cost more than planning assignments to that inside [10]". The aggregate execution time of work flows is frequently influenced by different latencies, for example, the asset disclosure, booking and information get to latencies for the individual work flow forms. To build information region and to lessen above latencies, Meta Data Service (MDS) is executed to store sets of datasets and their area on the proposed framework with the assistance of throttling system. The decision of capacity design [7] likewise significantly affects work flow execution time and the cost of nearly takes after execution, and so forth. This framework utilizes nearby reserve for information stockpiling.

Whatever remains of this paper is sorted out as takes after. Area 2 gives a review of the related work. Area 3 displays the proposed framework and its segments. Segment 4 reports the exploratory outcomes. Area 5 closes with a conclusion.

RELATED WORK

A workflow could be a high-level specification of a collection of jobs that represent procedure science or business flows and therefore the dependencies between the jobs that has got to be glad so as to accomplish a particular goal. The business progress is sometimes management flow-driven and includes constructs to specify methods, conditions, and should embrace human interaction. Typically, a business workflow implements a company's product or service. Scientific workflows usually touch upon massive an outsized quantity of knowledge and advanced calculations and so utilize large storage capacities and computing resources. During a scientific progress, a task is associate degree workable program with a collection of input parameters and files. Scientific workflows are generally information flow-driven and don't have made management flow structures, whereas notable exceptions exist, like Askalon [1].

In several analysis fields, particularly in lay to rest disciplinary field like bioinformatics and climate simulation, scientific workflows are typically each computation and data-intensive. Workflow running typically wants large-scale computing resources however additionally large storage. Today, scientific workflows applications carries with it many thousands of jobs, consume gigabytes or terabytes input information sets and generate similar amounts of intermediate data. These applications are noted as data-intensive progress applications. A knowledge intensive application consists of applications that manufacture, manipulate, or analyze information within the vary of many megabytes (MB) to petabytes (PB) [2]. Having the power [3] to makeover workflows between execution resources needs a precise degree of flexibility once it involves information placement and information movement. Most workflows need a knowledge storage resource near the execution web site so as to execute in associate degree economical manner. The benefits of cost-effectiveness, on-demand resource provision and straightforward for sharing.

Many models have been proposed and actualized to successfully execute the work process job. Different heuristic and meta heuristic techniques have been explored streamlining single or composite parameters for work process booking. An auto scaling technique to enhance the cost while meeting the due dates is proposed by Mao et.al.[6] proposed a Deadline Early Tree heuristic calculation with due date as limitation which limits cost. Rizos et al. [7] proposed a best moderate task approach LOOS and GAIN for improving the cost and execution time. The underlying assignments are made utilizing heuristic calculation which are reassigned to meet the financial plan and time requirements. Yu et.al [8] proposed a Markov Choice Process based way to deal with limit the cost of executing application on matrix supporting the client characterized due date. Zheng et al.[9] stretched out the HEFT calculation to Financial plan obliged Heterogeneous Earliest Finish Time (HEFT). The calculation considers the spending imperative and streamlines the execution time. Durillo et.al [10] examined multi objective HEFTI (MOHEFT) based rundown

booking heuristic which streamlines makespan and cost of executing work processes in Amazon EC2.

Over the prior time, the planning of information concentrated work flows is emerged as a critical research theme in conveyed figuring. Work flow planning for cloud is not the same as that in multiprocessors or lattice framework. This area concentrates on works managing information territory and information exchanges or information development in cloud condition. PiotrBryk et al. [6] proposed a novel element planning calculation that knows about the fundamental stockpiling framework that can be utilized on IaaS mists by considering information exchanges. This calculation upgrades plans by exploiting information reserving and document territory to lessen the quantity of record exchanges amid execution. Ke Wang et al. [8] proposed calculation for work taking (load adjusting) to decrease information area and information exchanging overhead to run information serious logical applications in Many-undertaking figuring

(MTC). Claudia Szabo et al. [9] proposed structure which incorporate portion and requesting chromosome for distribution of undertakings to hubs and execution arrange as per the logical work flow portrayal by considering information exchange and execution time utilizing hybrid and change administrators with the standard NSGA-II calculation. D.Yuan et al. [10] proposed k-implies bunching procedure for information position of logical Cloud work flows to diminish information development. Mingjun Wang et al. [11] additionally utilized k-implies bunching calculation to disperse input informational indexes into various server farms with the most related datasets set together. In this paper creators likewise proposed a multilevel assignment replication planning system to diminish enormous informational collections move in the runtime of the logical work flows.

Optimization of performance metrics such as quantity and potential in mixed computing situation develop sex trastimulating outstanding to the variance in the computing capability of accomplishment nodes and differences in the information transmission competency of communication associations among these nodes.

SaimaGulzar Ahmad et al. [12] obtainable a dual impartial Dividing based Data-intensive Workflow optimization Algorithm (DDWA) for heterogeneous calculating classifications. In this algorithm, the application job graph is divided so that the inter-partition information movement is minimal. PDWA provides significantly reduced latency with increase in the quantity.

THE PROPOSED SYSTEM

This section discusses the planned data-intensive workflow programming system. This method has 2 phases. The first section, tasks in workflow application square measure clustered so as to cut back execution overhead. the second section, these clustered workflows square measure allotted onto cloud resources like virtual machine (VM) supported the info neighbourhood as this planned system concentrate on data-aware programming. Within the data-intensive scientific workflows, tasks would like quite one dataset to execute. However, once these tasks square measure dead in several data centre, information transfer would become inevitable. To deal with these issues, this paper proposes clump methodology supported Task Dependency to optimize progress programming. The planned system conjointly uses Meta information Service (MDS) and information suffocation technique. By victimisation these techniques, the quantity of entomb cloud transfers and consequently the info traffic prices square measure heavily reduced and conjointly makes pan. The planned system for data-intensive progress programming is shown in [Fig. 2] and therefore the followings square measure parts of this planned system.

Application model

Technical workflows are showed as Directed Acyclic Graph (DAG). A DAG, G (V, E), contains of a set of vertices V, and edges, E. The edges characterise preference constrictions: each edge
$e_{a,b}$ = (t_a,t_b)∈Erepresents a preference constriction that directs that jobt_a should complete its accomplishment before task t_b $_b$starts. $e_{a,b}$also characterizes the volume of inter-task communication elaborate, e.g., the quantity of information(in bytes)that job$t_a$must direct to job$t_b$in order for job$t_b$to start its execution as shown in.

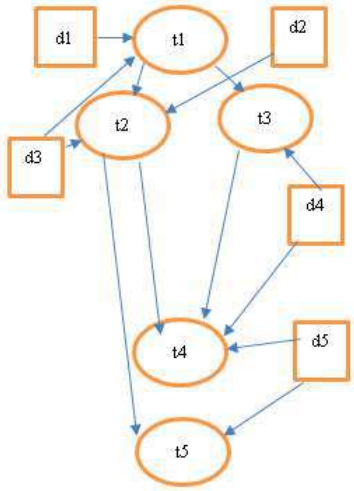


Fig. 1: Workflow diagram

Task grouping based on task dependency

Task grouping is a method that combines fined-grained jobs into coarse-grained jobs. Task grouping has verified to be an effective technique to decrease execution overhead and to increase the computational granularity of technical workflow jobs accomplishing on distributed resources. With job grouping, execution overhead can be removed by grouping small jobs are grouped together as one executable unit. This decrease benefits the cloud environment as an entire by decreasing traffic among the sites. The suggested scheme uses job grouping for fined grained tasks in user submitting workflow application to decrease information transfer time. First, we compute the dependences of tasks in workflow rendering to conditional probability.

$$P(T_a|T_b)=P(T_a \cap T_b)/P(T_a),\ P(T\square) > 0 \quad (1)$$

Which is the probability of task $t_a$occurring, given that task $t_b$occurs based on common usage of datasets. Clustering Method based on Task Dependency, noted as GTTD can be described as follows:

Algorithm 1.GroupingTechnique based Task Dependency (GTTD)

Input:task list task List
Output:all tasks are clustered
1. **Begin**
2. For $t_a \in$taskList Do
3. Calculate $P(t_a|t_b)$ for each pair of tasks
4. DM [a] [b] = $P(t_a/t_b)$
5. End For
6. Index=0
7. For each element ele(a,b) ∈DM Do
8. Selectmax (ele(a,b)) and a≠b
9. Marka and note Task $t_a$is most dependent on t_b
10. Add$t_a$and $t_b$to CL_{index} and remove row a and columna from DM
11. Index=index+1
12. End For
13. **End**

In our illustration, we use the example scientific workflow as shown in [Fig. 1]. In this figure, there are five tasksand five datasets. As flow described in GTTD, conditional probability matrix should be built from joint and marginal probability of each task pair in workflow. Joint and marginal probability table can be created from contingency table as shown in [Table 1].

Table 1: contingency table

t_b / t_a	T1	T2	T3	T4	T5	total
T1	1	1	0	0	0	2
T2	1	3	1	1	0	6
T3	0	1	2	2	0	5
T4	0	1	2	2	0	5
T5	0	0	0	0	1	1
total	2	6	5	5	1	19

Each value in contingency table is frequency of common usage of datasets for each task. According to Table 1, joint and marginal probability table can be built as displayed in [Table 2].

Table 2: Joint and marginal

t_b / t_a	T1	T2	T3	T4	T5	total
T1	0.090	0.090	0	0	0	0.180
T2	0.090	0.136	0.045	0.045	0	0.316
T3	0	0.045	0.090	0.090	0	0.225
T4	0	0.045	0.090	0.090	0	0.225
T5	0	0	0	0	0.045	0.045
total	0.180	0.316	0.225	0.225	0.045	0.991

Table 3: Conditional Probability

t_b / t_a	T1	T2	T3	T4	T5
T1	0.5	0.5	0	0	0
T2	0.284	0.430	0.142	0.142	0
T3	0	0.2	0.4	0.4	0
T4	0	0.2	0.4	0.4	0
T5	0	0	0	0	1

From [Table 2], conditional probability $P(t_a|t_b)$ for individually pair (t_a, t_b)of jobs can be designed as offered in [Table 3]. Constantly, job dependency for individually pair can be considered by means of conditional probability. In GTTD, jobs are grouped by the extreme value of to eachjob dependency from row by row if these pair of tasks is not same in number. In this case, if two pairs of jobs are not same in number, eg. $t_a \neq t_b$but concentrated value of tasks have same value; this techniqueprocedures the first pair of jobs as dependence. Task

Table 4: Task grouping by maximum conditional probability

t_b / t_a	t1,t2	t3	t4	t5
t1,t2	0.5	0	0	0
t3	0	0.4	0.4	0
t4	0	0.4	0.4	0
t5	0	0	0	1

(b)

t_b / t_a	t1,t2	t3,t4	t5
t1,t2	0.5	0	0
t3,t4	0	0.4	0
t5	0	0	1

After grouping, the projected scheme assigns jobs to resources in cloud environment with regard to data-aware development in [13]. Jobs are allocated to resources where essential input data are existed.

Meta information Service (MIS)

Data transfers among knowledge centres are unavoidable once running data-intensive advancement applications. The information transfer times don't seem to be negligible and should take a significant portion of the whole advancement execution time. Hence, each cloud system used for information intensive applications ought to showing intelligence assign the information to confirm a low level of worldwide data transmission quantity. This technique assumes that knowledge needed by the task should be out there at native cache of resources before each task will initiate execution and whereas running. It additionally assumes that knowledge are required for an advancement application is in a position to move

from one execution host to different. To cut back information access latencies and increase knowledge neighbourhood, the planned

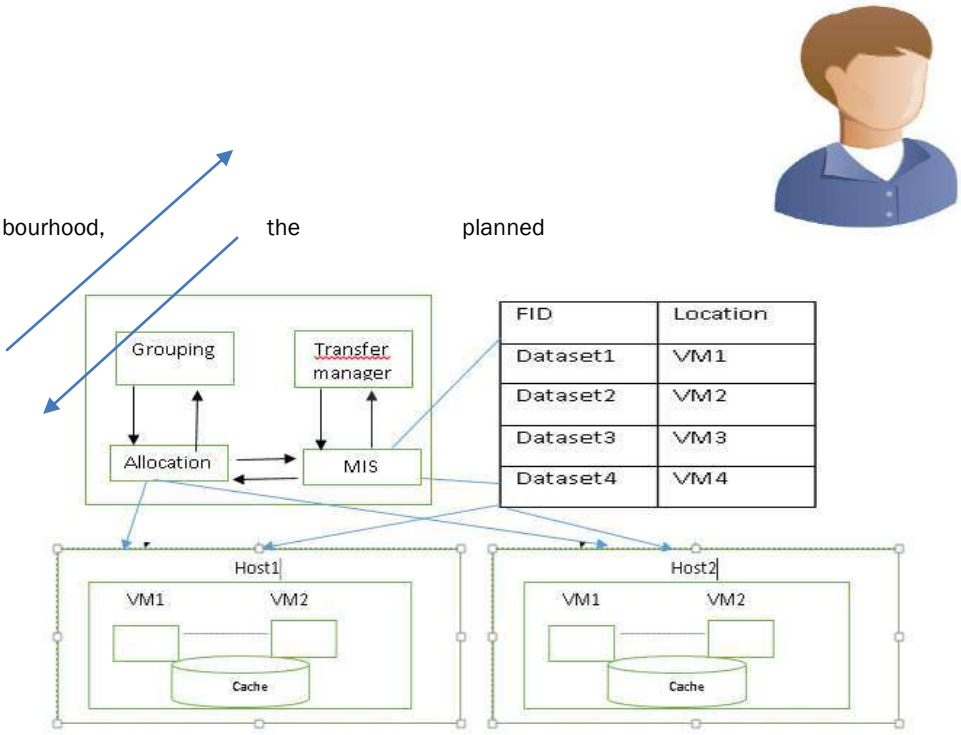


Fig. 2: Proposed system architecture

System uses MIS holds key-value sets: file-ids (FID) and site (locations of file) in many VMs. The successor tasks in another VM will begin execution when needed input file from their predecessor tasks is accessible. These successor tasks got to request to Transfer Manager to induce the placement of intermediate knowledge that are generated in runtime of execution. During this case, Transfer Manager will facilitate to search out location wherever knowledge will exist via MIS, as pictured knowledge asphyxiation.

In scientific workflows with data-intensive work, individual jobs might have to be required to watch for massive amounts of knowledge to be delivered or made by different tasks. Advancement systems arrange to succeed high performance by showing intelligence planning tasks on resources, with making an attempt to move the most important knowledge files on the highest-capacity links. If the network link has unrestricted information measure, knowledge asphyxiation moves knowledge to the tasks that desires the information a lot of desperately than the opposite. Knowledge asphyxiation will scale back the unnecessarily wasted information measure that will be utilized by the opposite applications if the congestion limits the information measure for a few transfers.

In the planned system, Transfer Manager uses knowledge asphyxiation technique to move massive size of knowledge on highest capability links. This part will facilitate moving largest knowledge files by asphyxiation up and down between links once it receives requests from the tasks. It's hoped that this system can scale back transfer time and additionally improve performance.

RESULTS AND DISCUSSION

In this research, Montage workflow application is chosen from dissimilar technical regions as this scheme focuses on information-intensive workflow. According to [14], Montage workflow has low CPU consumption for several job types in the workflow (mBackground, mImgtbl, mAdd, mShrink). This is as Montage jobs spend much of their time on I/O operations

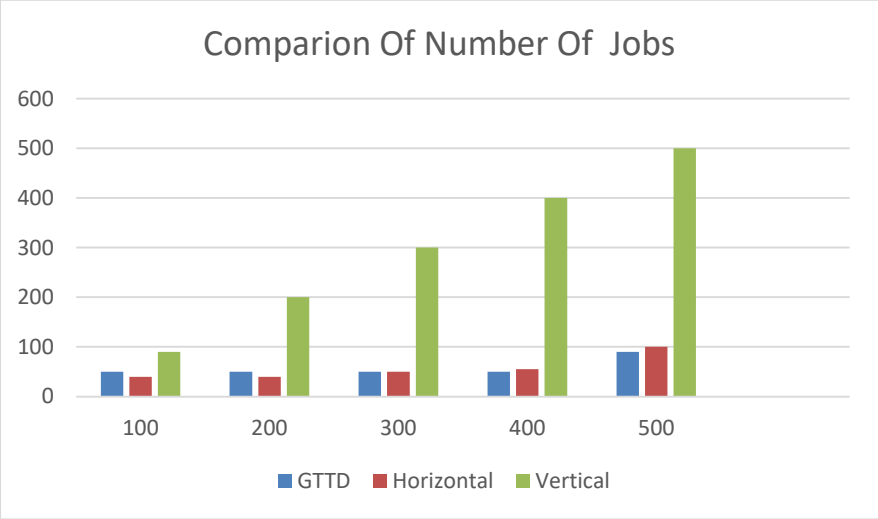


Fig. 3: Comparison of three Grouping methods.

For calculating the concert of the above grouping algorithm, WorkflowSim [13] framework is used in simulation. There are 20 single similar core virtual machines (worker nodes) are used to compose replicated computing platform. Each virtual VM has 2 cores as processing element (PE) and ability to process 1,000 million commands per second (MCPS). And network bandwidth is 15MB and 512 MB of memory. And Workflow Generator [15] is used to produce mock workflows through task runtimes and different number of tasks based on allocations gathered from running actual workflows.

CONCLUSION

Cloud computing has grown up in admiration in research community for organizing workflows. Tasks and datasets of workflow submission across worldwide circulated data centres need to be programmed conversing to information layout and makes pan for workflow implementation. This paper compares the quantity of jobs with additional grouping methods. The experimental consequence specifies GTTD beats others. Allowing to results, this suggested scheme will use GTTD to decrease execution overhead. It is predictable that this scheme will confidently aid the implementation of technical workflows well.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Ostermann S, Plankensteiner K, Prodan R, Fahringer T, and Iosup A. [2009]Workflow monitoring and analysis tool for ASKALON. In Grid and Services Evolution.

[2] Suraj Pandey, [2010] Scheduling and Management of Data Intensive Application Workflows in Grid and Cloud Computing Environments, PhD Thesis, December 2010.

[3] Deelman E, Vahi K, Juve G, Rynge M, Callaghan S, Maechling PJ[2014]Pegasus: a Workflow Management System for Science Automation, Future Generation Computer Systems, pp. 17-35

[4] Park SM Humphrey. [2008] Data throttling for data-intensive workflows. In: Proceedings of IEEE international symposium on parallel and distributed processing, IEEE, pp 1–11

[5] JigneshLakhani, Hitesh Bheda.[2012] Scheduling Technique of Data Intensive Application Workflows in Cloud Computing, NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE,

[6] Piotr Bryk, Maciej Malawski, Gideon Juve, Ewa Deelman. [2015] Storage-aware Algorithms for Scheduling of Workflow Ensembles in Clouds, Journal of Grid Computing, to appear

[7] Fuhui Wu, Qingbo Wu, Yusong Tan. [2015] Workflow scheduling in cloud: a survey, The Journal of Supercomputing.

[8] Wang Ke, Qiao Kan, Iman Sadooghi, Xiaobing Zhou, Tonglin Li, Michael Lang, Ioan Raicu.[2015] Load-balanced and locality-aware scheduling for data-intensive workloads at extreme scales, Concurrency and Computation: Practice and Experience, 00:1-29

[9] Zhengxiong Hou, Jing Tie, Xingshe Zhou I. Foster M. Wilde. [2009] ADEM: Automating Deployment and Management of ApplicationSoftware on the Open Science Grid. GRID 2009.

[10] Dong Yuan, YY Xiao liu, Jinjun Chen, [2010]A data placement strategy in scientific cloud workflows, Future Generation Computer System,26(8): 1200-1214

[11] Mingjun Wang, Jinghui Zhang, Fang Dong, JunzhouLuo, [2014]Data Placement and Task Scheduling Optimization for Data Intensive Scientific Workflow in Multiple Data Centers Environment, Second International Conference on Advanced Cloud and Big Data,

[12] SaimaGulzar Ahmad, Chee Sun Liew, M. Mustafa Rafique, EhsanUllahMunir, Samee U. Khan. [2014]. Data-Intensive

Workflow Optimization based on Application Task Graph Partitioning in Heterogeneous Computing Systems. IEEE International Conference on Big Data and Cloud Computing (BdCloud2014)

[13] Chen W and Deelman E.[2012] WorkflowSim: A Toolkit for Simulating Scientific Workflows in Distributed Environments, in The 8th IEEE International Conference on eScience

[14] Juve G, Chervenak A, Deelman E, Bharathi S, Mehta G, Vahi K.[2013] Characterizing and Profiling Scientific Workflows, Future Generation Computer Systems, 29(3): 682-692

[15] Workflow Generator, https:// conflence.pegasus.isi.edu/ display/Pegasus/Workflow Generator.

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

ENHANCED HADOOP PERFORMANCE ANALYSIS USING HADOOP ECO-SYSTEM

Shivam Suryawanshi[1], Jabanjalin Hilda[2*]

[1]*School of Computer Science and Engineering, Vellore Institute of Technology - Vellore, T.N, INDIA*

[2]*Faculty of School of Computer Science and Engineering, Vellore Institute of Technology - Vellore, T.N, INDIA*

ABSTRACT

All over the World, the idea of Big Data analytics is constantly growing in the software domain. Hadoop as an open source has become a popular Java framework for large scale data processing in recent years. Day by day, the rate of data is also increasing through different sources, and to analyze such huge data it is required to be processed on Hadoop Distributed File System (HDFS) rather than processing using traditional methods. The Apache Hadoop eco-system supports many open source tools for analyzing substantial datasets. Three well-known tools for data analyzing in the HDFS are Hive, Pig, and MapReduce. In this project, we are analyzing crime dataset which is imported on HDFS using Sqoop tool and further, it is analyzed using Hive, Pig, and MapReduce program. The result shows that it is an enhanced Hadoop performance analysis though Hadoop eco-system components where Hive works more efficiently than Pig and MapReduce Program. Later, results are compared with MapReduce program. In this crime data analysis, Hive outperforms 2.23 times than MapReduce and 1.90 times than Pig.

KEY WORDS
HDFS, Hadoop, MapReduce, Hive, Pig, Sqoop

INTRODUCTION

These days a lot of information are coming from various sources like Social network profiles, advanced media - audio, photos, and web sources, so on. The successful storage, querying and analyzing of these information has turned into a challenging test to the business. When it comes to crime, size of crime data growing more day by day and storing, analyzing, processing such large data has dependably been a big concern in the field of database management domain. Nowadays big data has become a part of handling such issues related to large information. There are a few inquiries with regards to crime. Questions like - Is crime a serious issue where you live? What sorts of crimes happen frequently? Is the crime rate increasing or decreasing where you live? Do you think the world will be secure or riskier later on? [1] And many more. To solve this problem, it is required to analyze the crime related data firstly using big data technology. It is also important to characterize certain definitions that are identified with Big Data and Hadoop.

Big Data

Received: 4 June 2017
Accepted: 12 Aug 2017
Published: 23 Sept 2017

Big Data are huge-volume, high-speed, as well as huge-variety data resources that require new types of handling to ensure upgraded process enhancement [2]. Increased processing speed, storage limit, and systems administration have made information to develop in every one of the 4 measurements. The four characterizing qualities of Big Data- volume, variety, value and velocity - are incorporated for the better performance of data handling. There are distinctive methods for characterizing and comparing Big Data with the customary information, for example, information size, content, collection and processing. Huge information has been characterized as expansive data sets that can't be prepared using conventional handling strategies, for example, Relational Database Management Systems, in an average preparing time. So for that new technology comes under big data is nothing but Hadoop [3], [4].

Hadoop

Hadoop is the system which allows to store and process a huge amount of data across multiple computers connected in distributed environment. By using the Hadoop technology, we can scale up from single-node cluster to multi-node cluster of machines, each offering different storage capacity and computation. Hadoop can be used in a different application in order to process the data as the data is generating more and more information on a daily basis, and it is becoming very difficult to handle the data. The importance of big data technologies is providing more accurate analysis, which can be used for decision-making in any business process. There are different big data technologies such as operational Big Data which includes a system like MongoDB where data is primarily stored [5], [20]. These systems are supposed to take an advantage of new distributed computing systems that have created over the earlier decade which will run productively. This tends to information workloads significantly less demanding to manage, simple, quicker, and less costly to execute. Another Big data technology is, which uses the parallel environment such as MapReduce programming that provides analytical capabilities to review and complex examination analysis all of the type of the data. We can use MapReduce to scale up the single machine to multiple machines. It provides different method for mapping the information which is integral to the capacities gave by SQL.
Using mapping function, it takes the data from the huge dataset and distributes to multiple machines. [Fig. 1] Shows the Hadoop Framework which incorporates MapReduce layer and HDFS layer. A little Hadoop group incorporates a master part and different slave part. The framework consists of a DataNode, NameNode, Job Tracker and Task Tracker [6].

***Corresponding Author**
Email: jabanjalin.hilda@vit.ac.in
Tel.: +91-7598193077

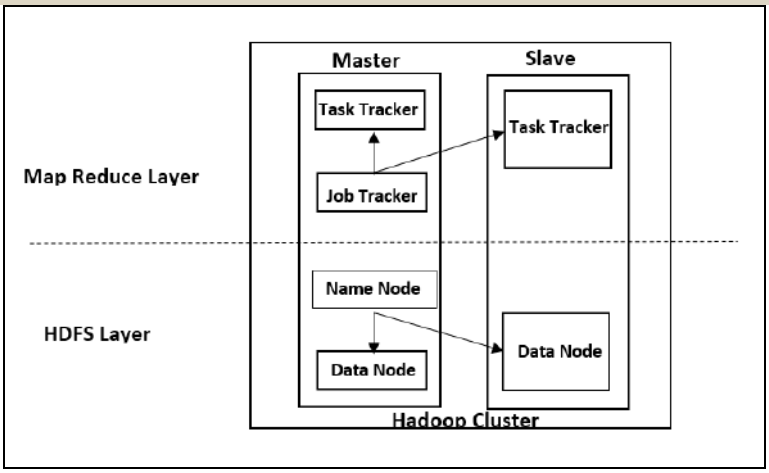


Fig. 1: Hadoop Framework [5]

- ***Data Node:*** Constantly inquire as to whether there is something for them to do, mainly used to track which data nodes are up or and which data nodes are down.
- ***Name node:*** Manages the record framework name space, it monitors where each block is present.
- ***Job tracker:*** Assigns the mapper job to undertaking tracker nodes that have the information or are near the information (same block)
- ***Task tracker:*** Keep the work as near the information as could be expected under the circumstances.

NameNode stores MetaData about the information being put away in DataNodes though the DataNode stores the real Data. JobTracker is an expert which makes and runs the jobs. JobTracker which can keep running on the NameNode distributes the job to TaskTrackers which keeps running on DataNodes; TaskTrackers run the assignments and report the status of the job to JobTracker. A slave part is responsible tracking job with the help of DataNode and TaskTracker. In a single-node cluster, both the NameNode and DataNode use the same machine for processing the data. In a multimode cluster, NameNode and DataNodes are ordinarily on various machines. There is one and only NameNode in a bunch and numerous DataNodes [6].

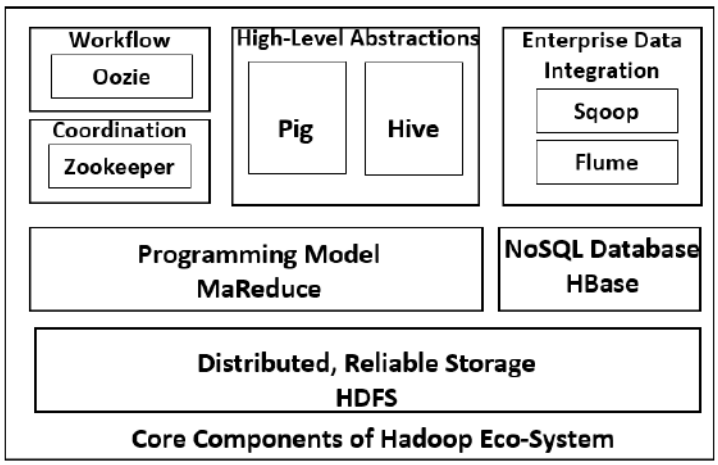


Fig. 2: Hadoop Eco-system components [8]

In addition, by enhancing the Hadoop we can reduce processing time, data size to read and other parameters in Hadoop MapReduce environment [7], [8]. This paper focuses on conceptual technologies, tools about huge information analysis and results shown through a couple of situations how it's useful for associations inside different segments if examination are conducted effectively. There are core components of the Hadoop eco-system, they are Sqoop, Pig, Hive, and Map Reduce. [Fig. 2] shows the architecture of Hadoop Eco-system components [9].

LITERATURE SURVEY

At current situation, each organization is confronting common difficulties which should be adapted up rapidly and effectively. Hadoop was incorporated with various segments that can be utilized for placing, executing and analyzing information significantly as well as proficiently. The decision of a specific utility relies on upon the necessities of the data analysis, the user technical knowledge, and the tradeoff between development time and execution time. Big data environment exhibits extraordinary turn for

organizations inside different parts to compete an upper hand. There are irregularities and difficulties inside Big Data analysis: adequate calculations to cover raw information or investigation, dependability and integrities of Big Data, information storage issues and MapReduce paradigm [5].

Fuad et al. [9], presented a conference paper which introduces the execution time of Hive, Pig, and MySQL Cluster on a basic information system with basic queries while the information is developing. In this, they have framed MySql database issue related to storing and processing information. The issue with MySQL Cluster is that as the information becomes bigger, amount of time to prepare the information increments and for that extra sources might be required. With Hadoop - Hive and Pig, handling time can be speedier than MySQL Cluster. Hence, three data analyzers with similar information model will run basic queries and to discover at what number of columns Hive or Pig is speedier than MySQL Cluster. They have worked on Group-Lens data set where an outcome shows that Hive is the most proper for this information model in a minimal effort hardware condition.

Prabhu et al. [10], presented a paper and they have taken web log information for the experiment and probed on Native Hadoop attributes that is a benchmark framework where they examined from the outcome i.e. when they streamline Hadoop framework attributes then they can enhance the framework execution. In this way they worked on enhancing the parameter in view of the framework assets and application and they also talked about why Hadoop setup must be transformed from its default to particular framework. After executing the experiment, they noticed that native execution has enhanced by 32.97%. For that they have considered couple of parameters.

Sathyadevan et al. [11], presented a paper where they explained about, crime data analysis and its prevention methods which can be a precise way of recognizing the common crime patterns. In their approach, they are predicting the areas where crimes are happening for more number of time and they can imagine crime inclined zones. In this paper they have used the idea of data mining where they are extracting the unknown existing features, valuable data from an unstructured node. Here they gone through an approach between software engineering and criminal equity to build up an information mining methodology that can help tackle crimes speedier. Rather than concentrating on reasons for crime event like criminal foundation of wrongdoer, political hatred etc., they are concentrating chiefly on crime variables of every day.

Alshammari et al. [12], presented a paper where they displayed Enhanced Hadoop (H2Hadoop) that permits a Name-Node to distinguish the blocks inside the cluster where particular data is present. In H2hadoop, input data is less, also input functions are lessened by the quantity of Data-Nodes conveying the main data-blocks that are again distinguished by sending a job to Task-Tracker. They have added some control feature on Name-Node whose purpose is to assign a task of particular data to a Data-Node without sending it to a whole cluster. They talked about the proposed work of H2Hadoop and showed the execution time of H2Hadoop which is efficient with respect to native Hadoop.

HADOOP ECOSYSTEM AND ITS METHODOLOGY

Problem Evaluation

With persistently expanding population, crimes and its rate dissecting related information is a tremendous issue for governments to settle on vital choices to keep up the peace. This is truly important to guard the residents of the nation from violations. The best place to admire opportunity to get better is the voluminous raw information that is created all the time from different sources by applying Big Data Analytics which breaks down to specific patterns that must be found, so that law can be kept up legitimately and there is a faith of security and prosperity among the people of the nation. In this paper, the three methodologies are differentiated with the help of a use case for Hadoop: Crime data investigation [13], [14].

The dataset examined in these tests were produced by a MapReduce program, using these crime data set as info, it is possible to check the output of the executed programs for precision. The issue is characterized further in the following segment, trailed by areas on the Hive, Pig and MapReduce solutions, and then the outcomes. Three well known tools for data analyzing occupant in the HDFS are Hive, Pig, and MapReduce. Hive gives a SQL like front end with a database foundation. Pig gives high level programming language to perform information processing that additionally empowers the users to misuse the parallelism innate in a Hadoop Cluster. MapReduce needs a PC program (frequently Java Programming) for inserting, handling and showing the output information. Hive and Pig produce MapReduce code to do the genuine performance analysis [15].

Sqoop

Sqoop is a command line tool used to transfer data from RDBMS to HDFS and vice versa [16]. Firstly, user can import data via sqoop from RDBMS (either MySql, SQL Server, PostgreSQL, etc.). After importing data

from RDBMS it will sink to HDFS using Hadoop MapReduce functionality. Following workflow shows Sqoop Architecture, [Fig. 3].

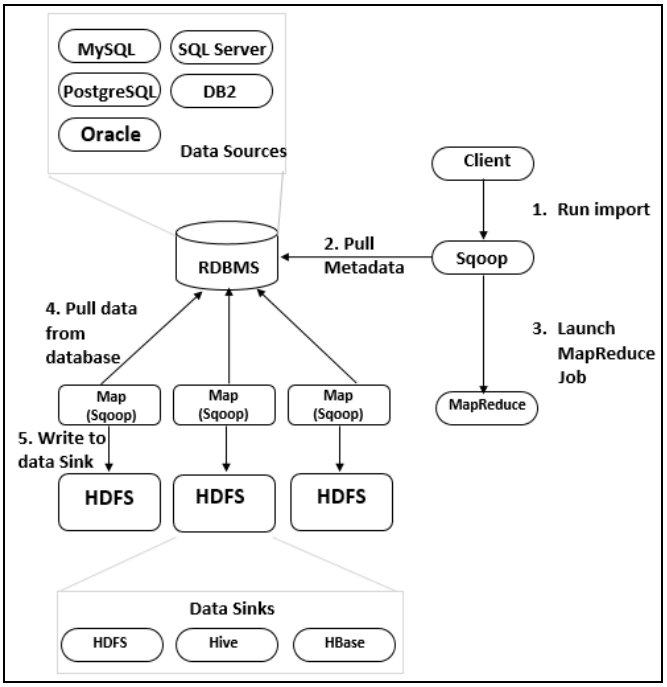


Fig. 3: Sqoop workflow architecture [16]

HIVE

The Apache Hive device is not a RDBMS tool, it is a part of the Hadoop eco-system, which works on the data which is stored in HDFS using HiveQL (HQL), is a Structured Query Language (SQL) interface, to solve or execute the query based on the available data [17]. This SQL based language in Hadoop domain gives good platform to view the information present in tables. Hive makes a query plan that implements the HQL in a progression of MapReduce projects, produces the code for these projects, after that executes the code, gives appropriate results. Following structure [Fig. 4] is the HIVE Workflow architecture. It shows that, how hive and Hadoop works together when it comes to MapReduce task.

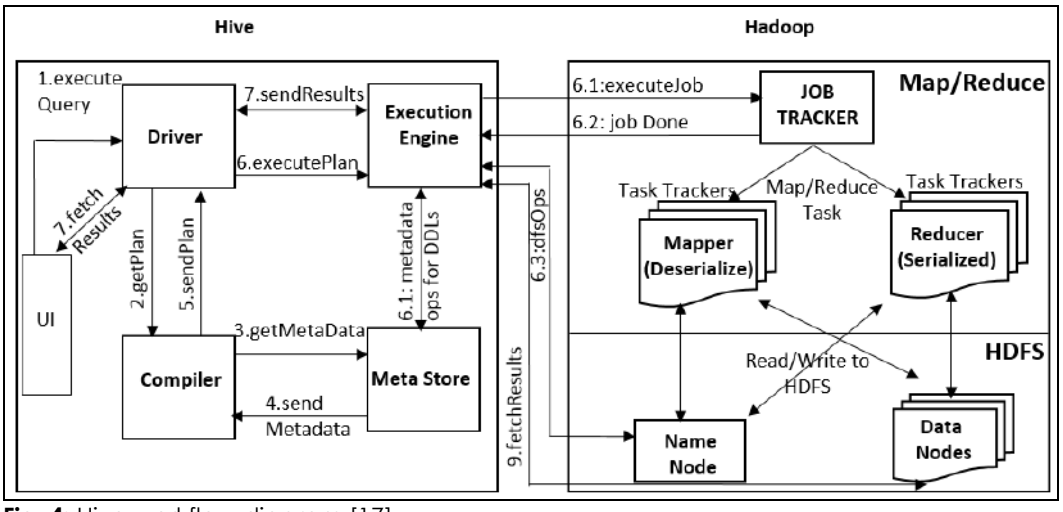


Fig. 4: Hive workflow diagram [17]

PIG

Apache Pig tool is another information analysis device in the Hadoop Eco-system [18]. Pig is a data flow language, Pig Latin, which allows the client to determine joins, and different calculations without the need to compose an entire MapReduce program. Like Hive, Pig creates a flow of MapReduce projects to solve the data analysis steps. Following flow chart [Fig. 5] will describe PIG architecture.

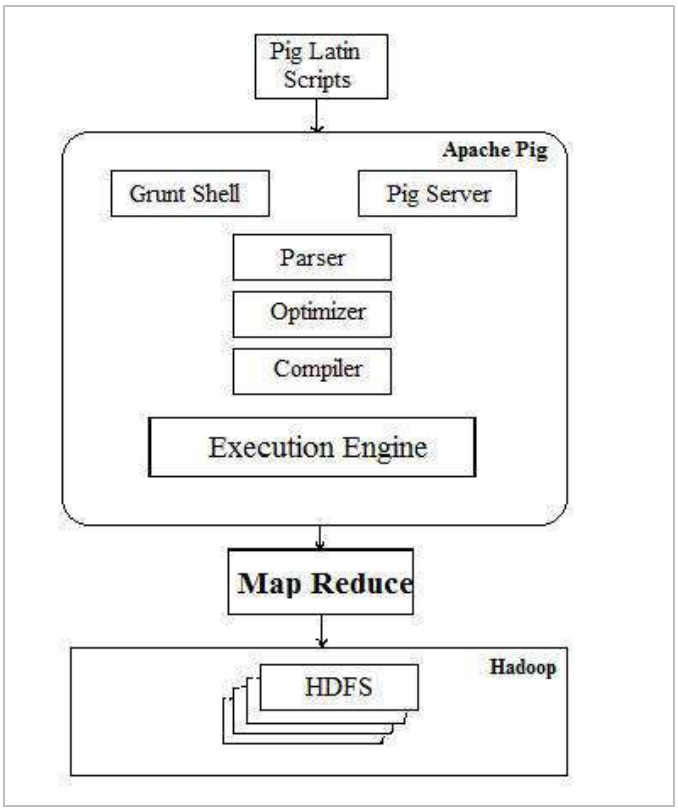


Fig. 5: Pig architecture [18]

Map reduce

MapReduce is a system utilized by Google for handling large measures of information in a distributed domain and Hadoop is Apache's open source execution of the MapReduce structure. Hadoop is helpful for putting vast volume of information into Hadoop Distributed File System and that information get prepared by MapReduce paradigm in parallel. MapReduce is a versatile and effective programming model to perform substantial scale information. At the point when handling this monstrous information asset has been constrained to single PCs, computational force and capacity rapidly get to be bottlenecks. This massive amount of information can be handled in distributed environment by processing each task one by one. The Hadoop MapReduce structure gives a stage to such parallelization of tasks [19].

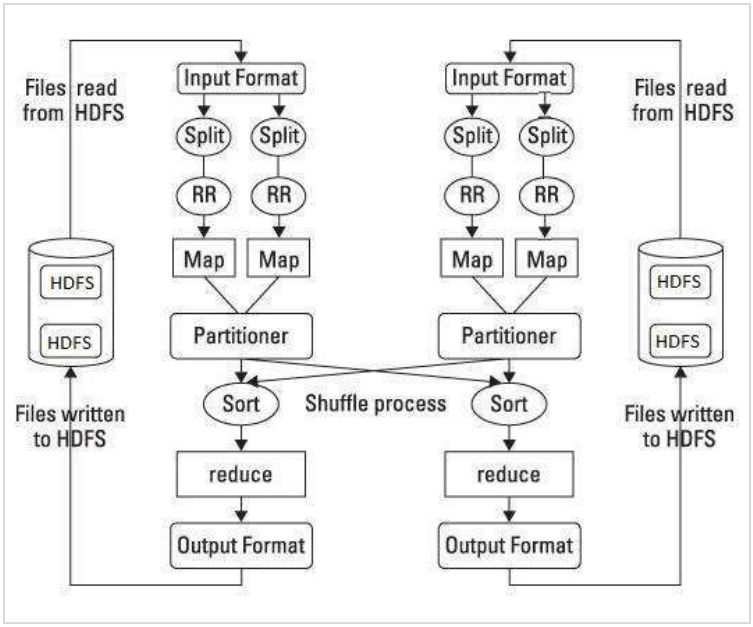


Fig. 6: Mapreduce paradigm. [19]

MapReduce Code Snippet

```
map(LongWritable k,Text v,Context c)
String s = v.toString();
        String s1[]=s.split(",");
        String type_crime=s1[6];
        c.write(new Text(type_crime), new IntWritable(1));
String report=s1[2];
        c.write(new Text(report), new IntWritable(1));
```

```
reduce(Text k,Iterable<IntWritable> v,Context c)
int count=0;
        while(v.iterator().hasNext())
                IntWritable i=v.iterator().next();
                count+=i.get();
        c.write(k,new IntWritable(count));
```

SIMULATION AND EXPERIMENTAL RESULTS

We have setup an experiment on Hadoop Cluster in Linux based Cent OS with single node running and configuration as: Minimum 2 GB RAM and Minimum 100GB Hard Disk Space. In this Crime Data Analysis, we are focusing on following three problem statements.

Table 1: Problem Definition

Problem #	Problem Statement	Result
Problem 1	Finding total number of crimes reported by	Fig. 7.1
Problem 2	Finding total number of Each Crime in last 6 Years (2011-2016)	Fig. 7.2
Problem 3	0.29 sec	Fig. 7.3

Using Sqoop Tool we are storing MySql data to HDFS. It can be done via following command. We can also revert the process i.e. HDFS to MySql [16].
Sqoop *->sqoop import --connect jdbc:mysql://localhost/training --username training --password training --table sr --m 1 --target-dir Crime_1;*
By using above statement, we can fetch the data present on the HDFS, Stored data further can be analyzed by hive and pig, results of each problem is summarized in the following section.

Hive analysis

Hive information is placed in HDFS, which is additionally sent to different nodes. This information is placed in a plain document with CSV, as provided by Sqoop. Hive will read entire file using indexing which results to faster query output. Hive won't execute a MapReduce Task if it does not include either of join, group by, order by aggregate operation. By querying this operation, hive can promptly begin the MapReduce task, which may requires 5-10 seconds to begin the MapReduce. [Table 2] shows Hive analysis. Solution using Hive for -

- ***Problem 1:*** *hive> select report_by,count(*)as tot from cdata group by report_by order by tot desc;*
- ***Problem 2:*** *hive> select" TOTAL CRIME FOR LAST 8 YEARS ", crime_type,count(*)as tot from cdata group by crime_type order by tot desc ;*
- ***Problem 3:*** *hive> hive> select year,count(*)as tot from cdata group by year order by tot desc;*

Table 2: Hive result analysis

Tool Hive:	Starting Time	Finishing TIme	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14.31:02	14.31.59	28.74	2	1	Successful
Problem 2	15:02:21	15:02:34	13.01	2	1	Successful
Problem 3	15.43:26	15:43:47	21.30	2	1	Successful

Pig analysis

Pig performs well with huge size of data. Pig executes a well ordered approach as characterized by the developer. If the query given by the developer is not a complex one (query which is included with joins and sorts) then Pig will not work properly. Pig solves every step one by one, which can expend more time in this case. Whenever information need composite job and more joining operation then Pig can deal with it productively by processing every level and persistently processing the next levels. Pig uses Grunt Shell to execute its task. [Table 3] shows pig analysis which is done after executing following script. Solution using Pig, for –

Problem 1: *grunt> cri1 = LOAD '/user/training/cri ' using PigStorage(',') AS (mon:int,year:int,report_by:chararray,loc:chararray,losa_code:chararray,losa_name:chararray,type_crime: chararray);*
grunt> cri2 = FOREACH cri1 GENERATE report_by;
grunt> by_loc = group cri2 by report_by;
grunt> count_crimes = foreach by_loc generate group as loc,COUNT(cri2) as TOT;
grant> cri6 = order count_crimes by TOT desc;
grunt> dump cri6;

Problem 2: *grunt> cri1 = LOAD '/user/training/cri ' using PigStorage(',') AS (mon:int,year:int,report_by:chararray,loc:chararray,losa_code:chararray,losa_name:chararray,type_crime: chararray);*
grunt> cri2 = FOREACH cri1 GENERATE type_crime;
grunt> by_type_crime = group cri2 by type_crime;
grunt> describe by_type_crime ;
grunt> count_crimes = foreach by_type_crime generate group as type_crime,COUNT(cri2) as TOT;
grant> cri6 = order count_crimes by TOT desc;
grunt> dump cri6;

Problem 3: *grunt> cri2 = FOREACH cri1 GENERATE year;*
grunt> cri3 = group cri2 BY year ;
grunt> cri5 = foreach cri3 generate group as year,COUNT(cri2.year) as TOT;
grunt> cri6 = order cri5 by TOT desc;
grant>dump cri6;

Table 3: Pig Result Analysis

Tool Pig:	Starting Time	Finishing TIme	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14.47:19	14.47.59	33	2	1	Successful
Problem 2	15:10:51	15:11:34	15.56	2	1	Successful
Problem 3	15.52:06	15:52:47	24	2	1	Successful

Map reduce analysis

Here CrimeDriver class contain Mapper and Reducer methods. After performing a following command, it will start execution where mapping, shuffling and reduce process will happen. [Table 4] shows result for three given problem statements using MapReduce Program. And [Fig. 6] gives MapReduce paradigm. In MapReduce scenario, when the database is compiled for the given problem statements 1,2,3 it requires 36,19 and 30 seconds respectively where number of mappers are 2 and reducer is 1.
[training@localhost workspace]$ hadoop jar crime.jar CrimeDriver cri MROUT1]

Table 4: MapReduce Result Analysis

Tool MapReduce:	Starting Time	Finishing Time	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14:22:14	14.22.50	36	2	1	Successful
Problem 2	15.20.24	15.20.43	19	2	1	Successful
Problem 3	15:30:42	15:31:12	30	2	1	Successful

The given problem statements are used to analyze the Crime Data and results of those are shown on above charts and graph using R tool. In first problem statement, we are finding total number of crimes reported by, result is shown in figure [Fig.7(1)]. Cambridge shire Constabulary 454435 (51.3%), City of London Police 41745 (4.3%), Durham Constabulary 390220(44%). In second problem statement, we are finding total number of Each Crime in last 6 Years (2011-2016), result is shown in figure [Fig. 7(2)], where Anti-social behavior 357187 counts more number of crimes from year 2001 to 2016. In third problem statement, we are finding total number of crimes per Year. In this case, total number of crimes are counted for year 2011 to 2016 in which crimes happened in year 2015 is close to 20000 which is again more. Result is shown in figure [7(3)]. Following part will show [Fig. 7(1)], [Fig. 7(2)], and [Fig. 7(3)] and also gives us an analysis using each of the Hadoop eco-system.

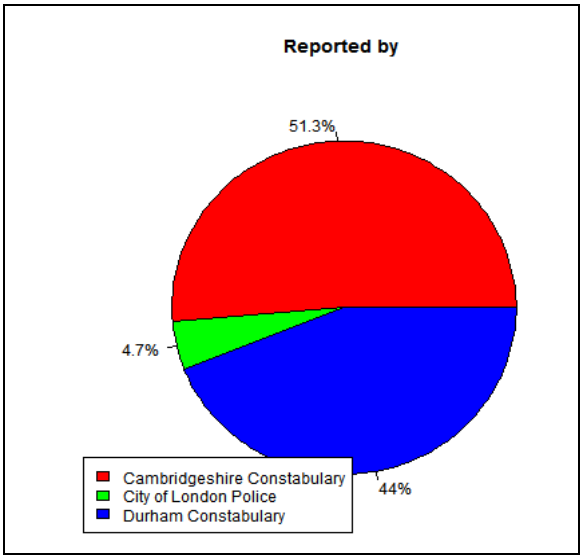


Fig. 7(1): Result for Problem #1

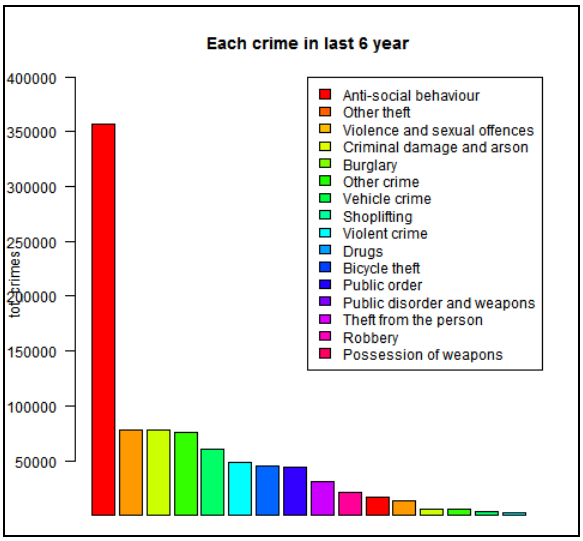


Fig. 7(2): Result for Problem #2

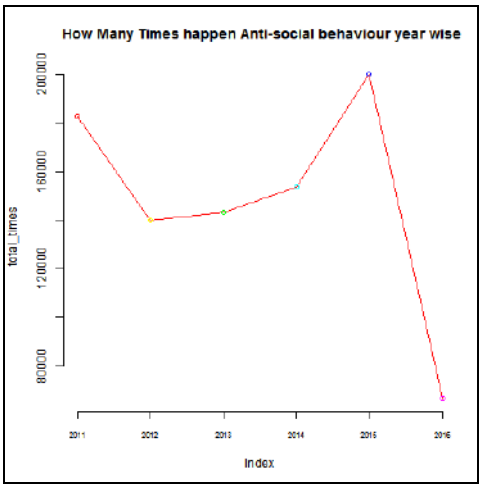


Fig. 7(3): Result for problem #3.

[Fig. 8] shows a graph where, we have analyzed the performance for each of the Hadoop Eco-system tool – Hive, Pig, MapReduce. Results of the graph shows that Hive works efficiently as compared to Pig and MapReduce for the given use case scenario: Crime Data Analysis. Similarly, we can analyze different problems related to crime data and results of this analysis which can be helpful for restricting crimes in future. If we consider average execution time for Hive, Pig and MapReduce, we can conclude that Hive is 2.23 times faster than MapReduce and 1.90 times than Pig for the given scenario. Also, average line of code used by Hive and Pig are very lesser than MapReduce Program. [Table 5] describes the comparison of each tool with respect to average time complexity and number of code lines.

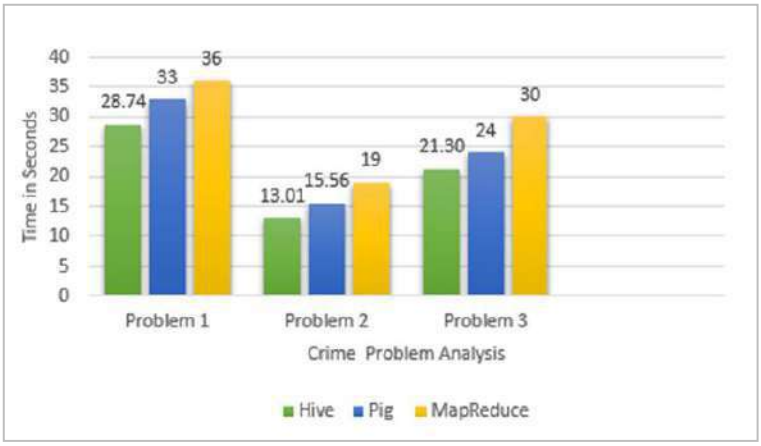


Fig. 7(4): Enhanced hadoop performance analysis.

Table 5: Hive, Pig, MapReduce Performance Comparison

Tool	Average Total Time Taken (seconds)	Time Relative to MapReduce	Line of Code Usage
Hive	38.05	2.23 times	2-4
Pig	72.56	1.17 times	10-12
MapReduce	85	1 .0 times	70-100

CONCLUSION

This paper presents data analysis of huge dataset utilizing three unique things that are a piece of the Hadoop environment – Hive, Pig and MapReduce. The application presented here is a crime Data investigation. The issue is clarified top to bottom and after the simulation, results are shown for the three tools. Complete dataset is accessible from https://data.police.uk/data/ and after successful operations, additionally the R techniques used to display the analysis and plot the outcomes. Results are appeared for each of the three tools with 8lacs set of records. Results show that Hive is more efficient when compared to Pig and MapReduce which shows that it is enhancing Hadoop Performance for huge data set. Hive is 2.23 times faster than MapReduce and 1.90 times than Pig. Also, line of code used by Hive and Pig are very lesser than MapReduce Program. In the future part, focus should be on finding the different parameters which can minimize the Hadoop performance for large size of data using different strategies.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
I express my sincere gratitude to Prof. Jabanjalin Hilda, Senior Assistant Professor in VIT University - Vellore, Department of Software Systems (SCOPE), for her expert guidance and invaluable support in Big Data – Hadoop Technology. I am extremely thankful to VIT University – Vellore for providing the infrastructure facilities for this project and also thankful to my Departmental Dean, HOD, all the Staff Members and my batch mates for their encouragement throughout the course of present work. Finally, I would like to acknowledge with gratitude, the support, and love of my family – my parents, friends without them it would not have been possible.

FINANCIAL DISCLOSURE
None

IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] "Introduction to Crime Data Analysis", Developed by Garner Clancey.

[2] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey", Journal of Big Data 2015.

[3] Pual C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, "Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill Osborne Media ©2011 book.

[4] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, et al., "Big data and its technical challenges," Communications of the ACM, vol. 57, pp. 86-94, 2014.

[5] S Vemula. "Hadoop Image Processing Framework."

[6] http://hadoop.apache.org

[7] T. White, Hadoop: The definitive guide: "O'Reilly Media, Inc.", 2012.

[8] Herodotou H.[2011] Hadoop performance models". arXiv preprint arXiv:1106.0940

[9] Ammar Fuad, Alva Erwin, Heru Purnomo Ipung, [2014]Processing Performance on Apache Pig, Apache Hive and MySQL Cluster", International Conference on Information, Communication Technology and System,

[10] Swathi Prabhu, Anisha P Rodrigues, Guru Prasad MS & Nagesh HR.[2015] Performance Enhancement of Hadoop MapReduce Framework for Analyzing BigData, Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on March 2015.

[11] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S, "Crime Analysis and Prediction Using Data Mining", Networks & Soft Computing (ICNSC), 2014 First International Conference on August 2014.

[12] Hamoud Alshammari, Jeongkyu Lee and Hassan Bajwa, "H2Hadoop: Improving Hadoop Performance using the Metadata of Related Jobs", IEEE TRANSACTIONS ON Cloud Computing 2015.

[13] Rachel Boba "Introductory Guide to Crime Analysis and Mapping", November 2001 Report to the Office of Community Oriented Policing Services.

[14] Arushi Jaina, Vishal Bhatnagara, [2015] Crime Data Analysis Using Pig with Hadoop", International Conference on Information Security & Privacy (ICISP2015), 11-12, Nagpur, INDIA

[15] Prachi Pandey, Sanjay Silakari, Uday Chourasia,[2016]A Comparative Study of Hadoop Family Tools", International Journal of Computer Science and Information Technologies, 7(3): 1620-1623

[16] http://sqoop.apache.org/docs/1.4.0-incubating/SqoopUserGuide.html

[17] http://hive.apache.org

[18] http://pig.apache.org

[19] Lu Jiamin, Feng Jun.[2015] A Survey of MapReduce based Parallel Processing Technologies", Big Data, Cloud & Mobile Computing,China Communication, Vol 11

[20] Yunquan Zhang, Ting Cao, Shigang Li, Xinhui Tian, Liang Yuan, Haipeng Jia, and Athanasios V. Vasilakos,[2016] Parallel Processing Systems for Big Data: A Survey, Proceedings of the IEEE 104(11).

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

PRIVACY PRESERVING IN DATA MINING USING DATA PERTURBATION AND CLASSIFICATION METHOD

Megha Dabhade[1], J Jabanjalin Hilda[2]*

[1]*School of Computational Intelligence and Engineering, Vellore Institute of Technology, Vellore, INDIA*

[2]*Faculty of school of Computational Intelligence and Engineering, Vellore Institute of Technology, Vellore, INDIA*

ABSTRACT

In today's information era, Tera bytes of data is being generated every second. It contains huge private and confidential information such as social networking, health care, finance, sensors data, criminal records etc. Data mining deals with such automatic generated data for different purposes. During such activities data gets exposed to other parties and protecting data becomes a challenge. The solution to this is provided by Privacy Preservation in Data Mining (PPDM). PPDM is the novel approach where different techniques are used to protect the privacy of data being used for Data Mining purpose. In this paper, we have done a comparative study on different data perturbation based privacy preserving methods and analyzed which one is more effective. In addition to this, Classification, the most commonly used Data Mining technique is used to develop a PPDM model. The experimental results demonstrate that how privacy is protected with respect to various privacy metrics.

KEY WORDS
PPDM, Data Mining, Classification, Data perturbation

Received: 11 June 2017
Accepted: 7 Aug 2017
Published: 23 Sept 2017

***Corresponding Author**
Email:
jabanjalin.hilda@vit.ac.in
Tel.: +91 7598193077

INTRODUCTION

Recent advances in communication, computing and digital technology data is growing incredibly day by day. Such data is distributed across geographical and administrative boundaries and it demands for a powerful technique to manage and analyze such huge amount of data, called Data Mining. Data mining is the process of extracting non-trivial and potentially useful knowledge or patterns from multiple large data sources [1]

Data mining systems deals with large amount of private and confidential data from social networking, healthcare, defense activities etc. This kind of information is non-sharable, protecting such data has become an important challenge and new research stream in the field of data mining. The concept of protecting confidential and sensitive information used during different data mining activities is termed as Privacy Preservation in Data Mining (PPDM) [1] [2]. PPDM is the new techniques which not only allows us to extract useful information and provide accurate results, but also helps to prevent loss of sensitive data [3].

Privacy preservation transforms dataset which contains confidential information into modified or altered form. Again, most importantly this information is hidden from unauthorized users. Privacy preserving data mining i.e. PPDM is the emerging area in data mining where efforts are being made to protect private information from unauthorized revelation. Privacy Preservation Data Mining [1] was introduced by keeping security of sensitive information as a priority in data mining process and to furnish canonical data mining process. A large portion of these security protection methodologies were proposed to safeguard private data of test dataset. Then again, protection safeguarding process which conceals data, may decrease utility of these altered dataset. At the point when their utility reduces to a specific level, the minimized data may lead to inaccurate analysis [1] .

This paper mainly focuses on how PPDM can be effectively used using Data Perturbation and Classification method. Data perturbation is widely used technique for privacy preservation. In this technique, data which is to be processed, is modified before passing to data mining process. There are different ways to modify data like data distortion, data swapping, noise addition, data hiding etc. [2] [3]. Among these method data distortion is proved to be most popular and effective method

Classification is the most commonly used data mining technique to build model. Its main objective is to build a classifier to identify class label of data based on training data [4][5]. Classifier can be represented by using decision trees, Naïve Bayes classifier, Neural Networks, SVM etc. In this paper, we mainly focus on issues related with PPDM using decision trees. Privacy preservation of individual data and accuracy of constructed classifiers examines the performance of privacy preserving technique [6][7].

In this paper, Decision tree is used as classification method. This is supervised learning. The complex decisions are further dived into smaller decisions. The complexity is controlled by the pruning technique. It can handle both numerical and categorical data. Decision trees are also easy to interpret. More precisely, ID3 algorithm is used as decision tree algorithm. For implementing such trees, ID3 is most efficient among machine learning approaches. In this approach tress are constructed based on entropy or information gain values. The original data set is divided into training data and testing data. The classifier is calculated based on training data and it is further used to predict the class for testing data [8][9].

THE IIOAB JOURNAL
COMPUTER SCIENCE: Guest editor-Prof. S. Prabu & Prof. Swarnalatha P

IIOAB JOURNAL
ISSN: 0976-3104

RELATED WORK

The privacy preservation in data mining can be done in two ways. In the first approach, data mining algorithms are modified without any knowledge of data. While in second approach, methods modify the datasets values to keep the privacy of data safe. In this paper, we are concentrating on second approach, many researches has been carried out in data perturbation. Some of them are as follows:

Santhosh et al. were the first to propose the idea of Data swapping in the year 1982. This is transformation technique where dataset is modified by altering the values of attributes of datasets from selected records. The data swapping is proved to be a distinguished data perturbation technique for privacy preservation [10].

Naga et al. in 2006 have proposed Singular Value Decomposition (SVD) approach which based on data distortion approach. They have used real world datasets for their experiment and proposed further innovation called sparsified SVD [11] [12]. The experimental results showed that this new sparsified SVD is efficient in preserving privacy and it also maintains the performance of the datasets.

Aldeen et al. in the year 2012, have proposed data distortion strategies viz. SVD and sparsified SVD along with feature selection. Their main objective was to reduce feature space in features. There are different kinds of privacy metrics which assess the utility of data. These measures calculate the performance of data mining processes by finding the difference between original dataset and distorted dataset and what is the degree of privacy protection. The real-world dataset was used for experiment and the results demonstrated a feasible solution with the use of sparsified SVD than only SVD. [13-16]

Mohammed et al In the year 2011, a study was carried out on intuitionistic i.e. a proof of contradiction method, for fuzzy clustering and application of fuzzy k-member clustering to protect privacy concretely in pattern recognition. K-member clustering is k-anonymity clustering technique in which data samples are summarized so that every sample is different from at least (k – 1) other samples. To improve quality of data summarization with k-anonymity, a fuzzy variation of k-member clustering was proposed. A secure framework was proposed for handling both vertically and horizontally distributed data in case of fuzzy co-clustering[17]

Kake et al. proposed Fuzzy based PPDM in the year 2016 where fuzzy-based mapping techniques were compared in terms of their privacy-preserving feature and their ability to employ exactly same relationship with other fields. [18] A fuzzy c-regression method was used to generate synthetic data on which statistical computations were done by third party. In fuzzy clustering approach. This method effective because it collects records into clusters where each record is not recognizable from others after within-cluster merging. Hence lossless data anonymization can be achieved.

METHOD

Classification

In this paper Decision tree is used as Classification method. In the field machine learning and statistics, the decision tree algorithm is called as “Predictive modelling technique” which build a simple tree to construct the pattern of classification data. Decision is being popular because it has ability to handle both numerical and categorical data [8] . As well they are easy to interpret. It is inverted directed tree having root at the top and has peculiarity that any complex decision making process can be converted into smaller and simple decision.

Data perturbation

In this paper, we are using data perturbation method for modifying data. Data perturbation has an important aspect in preserving the privacy of data. Perturbation is deviation of system from normal state to some other but consistent state . After perturbation, the original data set is modified and further given for the analysis process. Data perturbation can be done in several, however, the most widely used techniques are: probability distribution and data distortion. Data perturbation is easy and effective technique for preserving confidential data [6] [9]. A no. of methods has been proposed for privacy preserving in data mining. This paper mainly focusses on five major methods used for data perturbation. Those are as follows:

a) ***Noise addition:*** In this method, the origin data matrix is added by uniformly distributed noise matrix. The noise matrix is of same size as original. The elements of noise matrix are the randomly generated numbers collected from continuous uniform distribution.

b) ***PCA:*** The principle Component analysis is mainly used for dimensionality reduction. In PCA orthogonal transformation is used, so to transform the original data of co-related samples into the set of linearly uncorrelated samples. This sample is known as Principle Components. PCA becomes sensitive when

original variables show relative scaling in their values [9]. It is a widely-used tool in exploratory data analysis and developing predictive models for decision making. There are two major ways for performing PCA one by Eigen-value decomposition of a data matrix or singular value decomposition of a data matrix.

c) ***SVD:*** SVD i.e. Singular Value Decomposition is frequently used method for data perturbation. It is usually used for the dimensionality reduction of the original data set. In this paper, it is used for data perturbation method [6] [17]. Let say, A be the original matrix of order m x n. The row(n) in matrix represents the data whereas column(m) represents the attributes. The SVD of the matrix M is:

$$M = U \sum VT$$

Where U is an orthogonal matrix of order m x n, $\sum$ is an m x n diagonal matrix whose diagonal elements are positive and VT represents an n x n orthonormal matrix

d) ***QR:*** It is primarily used for the decomposition of a matrix. Modified matrix is a product of orthogonal matrix (Q) and upper triangular matrix (R). This can be represented as follows:

$$M = QR$$

If M is a complex square matrix, then there is a decomposition M = QR where Q is a unit matrix (i.e. Q*Q = I). If A has m linearly independent columns, then first m columns of Q form an orthonormal basis for the column space of A. In other words, the first k columns of Q form an orthonormal basis for any $1 \leq k \leq n$. In short any column k of A depends only on the first k columns of Q which is amenable for the triangular form of R [9] [19].

In this paper, we have come up with approach where we are using different data perturbation methods to protect or preserve the privacy of data used for data mining processes. The following figure shows the functional workflow for PPDM. The classification used to verify the performance of the original dataset after perturbation. The privacy measures are calculated for each of data perturbation methods described above. [Fig: 1] depicts the functional workflow of our implementation.

Workflow

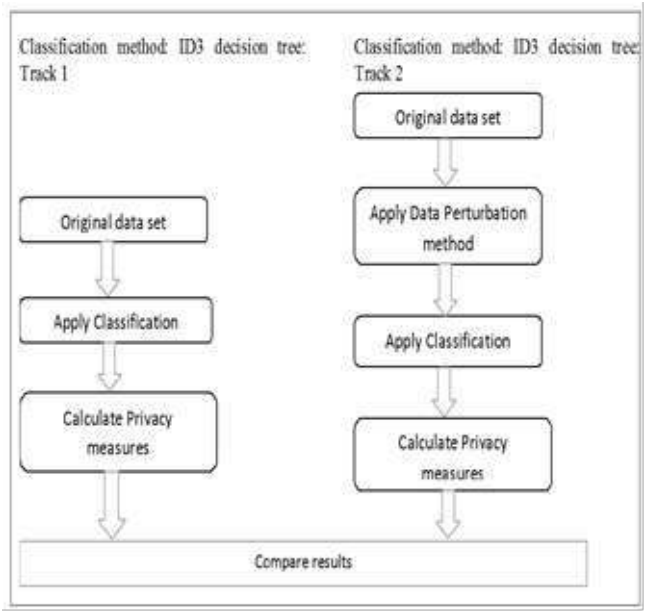


Fig. 1: Functional workflow.

The entire workflow of this paper is divided into main three modules viz:

a) ***Classification module:*** In this module, data set is being mined with one of the classification algorithm. In this paper, we have used "decision tree" as classification method. As well, accuracy is also calculated which being compared with perturbed data is set accuracy.

b) ***Perturbation module:*** In this module, the same data set is being perturbed using one of the perturbation method. We have used Noise addition as perturbation method for one of the column of dataset. We have also checked accuracy after perturbation as it will depict the distortion of data due to perturbation.

c) ***PPDM module:*** This is the important module where actual privacy parameters are being calculated which shows the results how perturbation method is useful for preserving the privacy of data. This module is applied over above two modules i.e. classification module and perturbation module.

IIOAB JOURNAL
ISSN: 0976-3104

PRIVACY MEASURES

In this paper, we have used privacy measure that are usually used by the PPDM methods, based on matrix decomposition . Privacy is said to be protected if VD, RP and CP have larger value and RK and CK will have smaller value.

a) ***Value Difference (VD):*** After applying perturbation on the data samples, the data gets modified. The modified changes are the value difference (VD) [1] between the original data and perturbed data. It is given by the relative value difference in Forbenius norm. The value difference is the ratio Forbenius norm on original data (A) and the perturbed data (PA) to the original data (A).

$$VD = |A-MA|_F / |A|_F$$

b) ***Position Difference (RP):*** After Data Perturbation on the dataset, the relative position of the data sample is modified also. There are many metrics to measure the positional difference of the data samples [1].

$$RP = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} | Rank_j^i - MRank_j^i |}{nm}$$

c) ***RK:*** It exhibits the percentage of elements that keep their values in each column after distortion. It is used to represent the average change of order for every attribute in data sample. After the data of an attribute is perturbed, the order of each data is changed. Let us say original data A has n observation and m attributes. Orderij depicts the ascending order of the perturbed sample Aij. The RK is defined as:

$$RK = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} | Rk_j^i}{nm}$$

Where, RK gives whether a sample retains its position in the order of the value:

$$Rk_j^i = \begin{cases} 1 & Rank_j^i - MRank_j^i \\ 0 & otherwise \end{cases}$$

d) ***CK:*** Like RK, CK can be defined to calculate the percentage of the attributes that retain their orders of average value after the perturbation. Hence CK is given as follows:

$$CK = \frac{\sum_{i=1}^{m} Ck_i}{nm}$$

Where CK_i is calculated as follows:

$$Ck_i = \begin{cases} 1 & Rank_j^i - MRank_j^i \\ 0 & otherwise \end{cases}$$

e) ***CP:*** The values of an attribute can be inferred from its relative value difference compared with the other attributes. Hence it is necessary to know the order of average value of the attributes that varies after the data perturbation. The CP metric can be used to define the change of the average value of attributes:

$$CP = \frac{\sum_{i=1}^{m} |Rank_j^i - MRank_j^i|}{nm}$$

Where RankVi, is the ascending order of the attribute, while MRankVi represents its ascending order after the perturbation[19].

The higher value of RP and CP and lower value of RK and CK, denotes the more privacy is preserved for given dataset.

IMPLEMENTATION

The idea introduced in this paper has been implemented with RStudio. It is the platform used for R language to develop programs especially in machine learning and data mining. R language is used to demonstrate classification method (Decision tree). [Fig: 2] shows decision tree with more specification like pruning which restricts unusual growth of tree and [Fig: 3] provides probability of each class in tree for better understanding of classifier.

IIOAB JOURNAL
ISSN: 0976-3104

Dataset: Cardiotocography Data Set
(https://archive.ics.uci.edu/ml/datasets/Cardiotocography)
Number of Instances: 2126
Number of Attributes: 23

EXPERIMENTAL RESULT AND ANALYSIS

Classification results

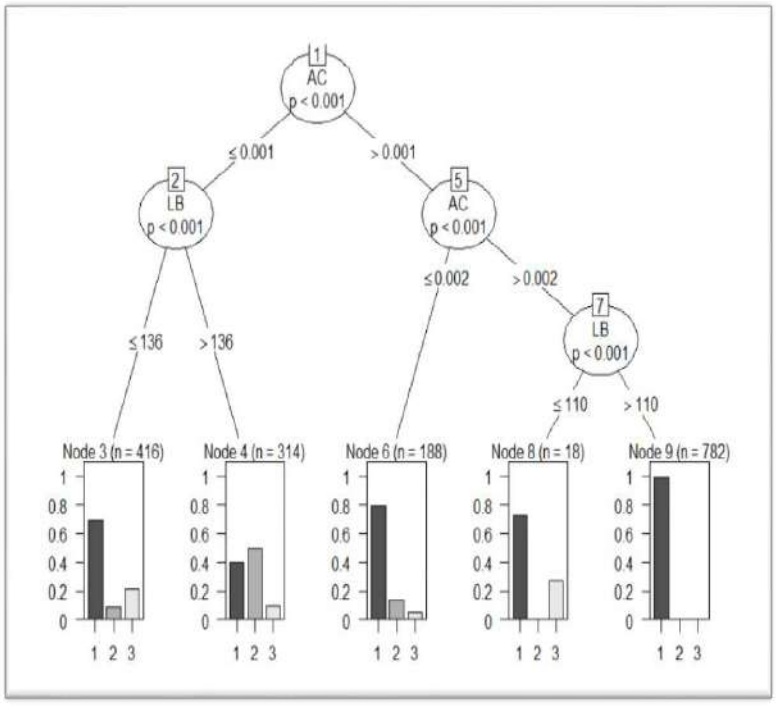


Fig. 2: Decision tree after pruning.

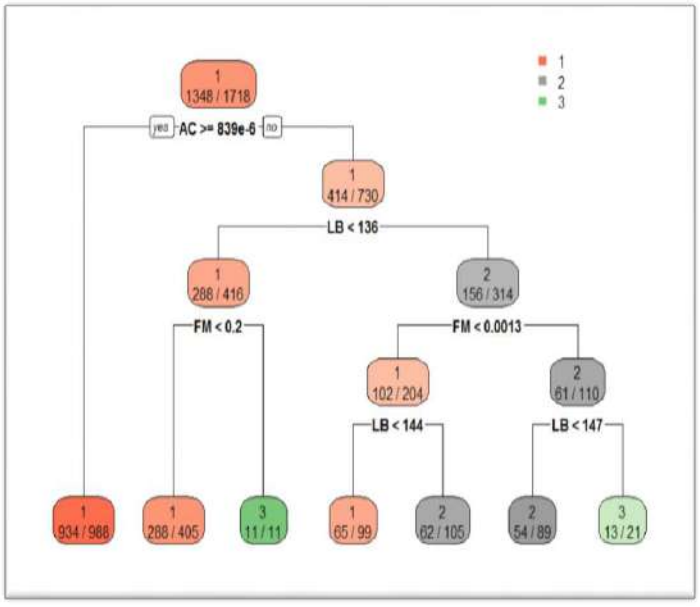


Fig. 3: Decision tree with probability of classes.

Classification Accuracy: It is the ratio of number of correct predictions to the total number of predictions made which is multiplied by 100 (%).
Again, only accuracy is not enough to conclude any classification or perturbation method to be efficient. Some other performance parameters must have calculated.

[Table 1] shows classification results while [Table 2] shows privacy measures calculated for four different data perturbation methods.

Table 1: Classification results

	Dataset without Perturbation	Dataset with Perturbation
No of Correctly Predicted classes	1389 (training data) 323 (testing data)	1430 (training data) 324 (testing data)
No. of wrongly Predicted Classes	329 (training data) 85 (testing data)	270 (training data) 84 (testing data)
Accuracy	0.7916667	0.7941176

PPDM results

Table 2: Privacy measures calculated for data perturbation methods

Privacy Measures	VD	RP	CP	CK	RK
PCA	1.000041	44.07045	2061263	1	0
SVD	1.000001	43.77193	2047301	1	0
Simple Noise Addition	0.006210513	0.0796302	3724.464	1	0.95455
QR	1.423345	45.09783	2109316	1	0

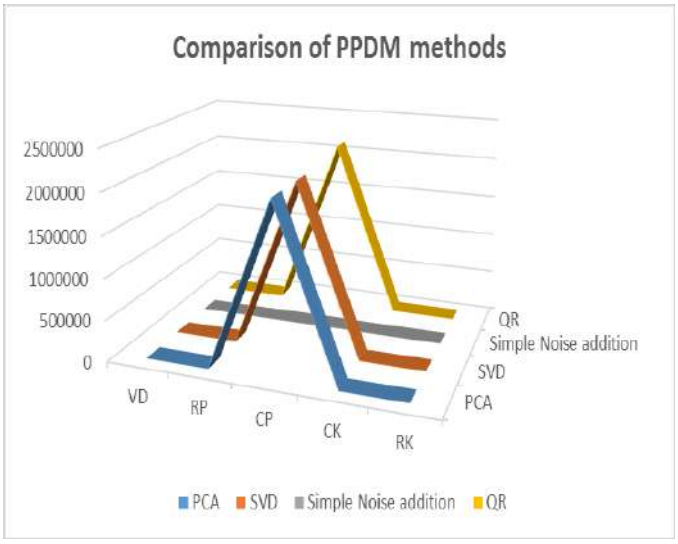


Fig. 4: Comparison between different data perturbation method.

Analysis

The experimental results show that PCA is the best among rest of three methods. Because it has higher value of RP and CP while lower values for CK and RK than other methods, which is considered as an efficient with regards to privacy measures. For simple noise addition method though it has the lowest VD, it can't be best because rest of measures are not significant for it to be best. Our focus is also on the privacy preservation classification; the experimental results clearly shows that classification is highly effective in terms of accuracy. The accuracy is getting preserved even after data perturbation hence both dataset used for experiment and data mining process is the best combination to perform out PPDM process.

CONCLUSION

This paper discussed various data perturbation methods and privacy measures have been calculated for each of the method. Hence Privacy Preservation in Data Mining i.e. PPDM proved to be novel approach to protect data. In addition to that Classification (Decision Tree) also played important role in efficient execution of Privacy Preserving methods and preserving accuracy of data. Though Data perturbation is quite obsolete method, its combination with Classification method made the approach very efficient. As

IIOAB JOURNAL
ISSN: 0976-3104

the field of Privacy is growing day by day, it important to establish a framework considering variety of privacy protecting algorithms.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof. Jabanjalin Hilda, Senior Assistant Professor, Department of Computational Intelligence, VIT University - Vellore, for her expert guidance and valuable support throughout project especially in Data mining. I am extremely thankful to VIT University – Vellore for providing me such great infrastructure facilities for this project as well I want to thank our Departmental Dean, HOD, all the other Staff Members and my batch mates for their encouragement throughout the course of project work. Lastly, I would like to acknowledge with gratitude, the support, and love of my family – my parents, friends without them it would not have been possible.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Samir Patel, Kiran R. Amin. [2013] Privacy Preserving Based on PCA Transformation Using Data Perturbation Technique, International Journal of Computer Science & Engineering Technology, ISSN: 2229-3345, 4(5): 477-484

[2] Inthumathi MS, Damodharan P.[2016] PPDM and Data Mining Technique Ensures Privacy and Security for Medical Text and Image Feature Extraction in E-Health Care System, International Journal of Computer Science and Information Technologies, 6 (6):5126-5129

[3] Rajesh N, Sujatha K, Arul Lawrence Selvakumar A. [2016] Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms, International Journal of Computer Applications (0975 – 8887) 133(7):20

[4] Suchitra Shelke, Babita Bhagat.[2015] Techniques for Privacy Preservation in Data Mining, International Journal of Engineering Research & Technology, 4(10)

[5] Bhupendra Kumar Pandya, Umesh Kumar Singh, Keerti Dixit.[2015] A Robust Privacy Preservation by Combination of Additive and Multiplicative Data Perturbation for Privacy Preserving Data Mining, International Journal of Computer Applications (0975 – 8887) 120(1)

[6] Yousra Abdul Alsahib S. Aldeen1, Mazleena Salleh and Mohammad Abdur Razzaque, [2015] A comprehensive review on privacy preserving data mining, Springer Plus 4:694 DOI 10.1186/s40064-015-1481

[7] Tamanna Kachwala, Sweta Parmar. [2014] An Approach for Preserving Privacy in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, 4(9)

[8] M.Syamala kumari, B.Govinda lakshmi, Privacy Preserving of Unrealised data sets using classification, IJCEA, Volume VII, Issue II, August 2014

[9] Sharmila A Harale, Bongale AK. [2014] Privacy Preservation and Restoration of Data Using Unrealized Data Sets, International Journal of Engineering Research and Applications, ISSN: 2248-9622, 4(7) :107-111

[10] Santosh Kumar Bhandare. [2013] Data Distortion Based Privacy Preserving Method for Data Mining System, International Journal of Emerging Trends & Technology in Computer Science, 2(3)

[11] Naga Lakshmi M, Sandhya Rani k.[2013] SVD based Data Transformation Methods for Privacy Preserving Clustering, ISSN: 0975 – 8887, 78 (3)

[12] Nagendra kumar.S, Aparn .R. [2013] Sensitive Attributes based Privacy Preserving in Data Mining using k-anonymity, International Journal of Computer Applications (0975 – 8887) 84(13)

[13] Aldeen YAAS, Salleh M, Razzaque MA.[2016] A comprehensive review on privacy preserving data mining, SpringerPlus, 4(1): 1-36

[14] Xinjun Qi, Mingkui Zong.[2013] An Overview of Privacy Preserving Data Mining, ICESE 2011, Procedia Environmental Sciences 12 (2012) 1341 – 1347 International Journal of Computer Applications (0975 – 8887) 8(13

[15] Guang Li, Yadong Wang.[2012] A Privacy preserving classification method based on single value decomposition, The International Arab Journal ofInformation Technology, 9(6)

[16] Mohammad Reza Keyvanpour, Somayyeh Seifi Moradi. [2011] Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework, International Journal on Computer Science and Engineering, ISSN: 0975-3397 (2) : 862- 871

[17] Kamakshi P, Vinaya Babu A.[2010] Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data, Journal of Computing, 2(4)

[18] Keke Chen, Ling Liu.[2009] Privacy-preserving Multiparty Collaborative Mining with Geometric Data Perturbation, IEEE Transaction on Parallel and Distributed Computing

[19] Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham. [2006]The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

REVIEW

A REVIEW OF CATEGORICAL DATA CLUSTERING METHODOLOGIES BASED ON RECENT STUDIES

N. Sowmiya[1], B. Valarmathi[2]*

[1]*School of Information Technology and Engineering, VIT University, INDIA*

[2]*Department of Software and Systems Engineering, School of Information Technology and Engineering, VIT University, INDIA*

ABSTRACT

In day to day activities, a very large volume of information is collected in all fields. The data mining task is necessary to handle those large amounts of data's. Clustering is the fundamental task in data mining, its main objective is to partition the dataset consists of 'p' objects into 'q' clusters. This paper presents the literature review of the clustering algorithm for categorical and binary attributes. Many algorithms were proposed in the literature for clustering categorical and binary data. The review is based on the type of methods used for clustering categorical data, evaluation criteria, datasets used, and input & output parameters. The objective of this review is to show which algorithm performs well when compared to the clustering accuracy obtained from various methods for similar datasets.

KEY WORDS
Data mining; Clustering; Categorical Data; Boolean values; K-means; Hierarchical;

Received: 28 May 2017
Accepted: 25 July 2017
Published: 24 Sept 2017

INTRODUCTION

In a real life senario a large a very large volume of information is collected in the field of medicine, academics, market basket data transactions, banking, and etc. To handle these large amount of data, data mining concept was evolved and still in the emerging area of research since 1960's. Data mining is an extraction of information from a large set of the database. Many data mining techniques are available for the extraction of knowledge. Some of the techniques include Classification, Clustering, Association Rule Mining, Prediction, and etc. Similarly, many algorithms were available for each technique. Our focus is on clustering technique. Clustering is used to group the similar objects together in one group and dissimilar objects in other group, the dataset is partitioned into 'q' clusters based on similarity or distance measures [58]. For good quality of the cluster, the inter-cluster similarity is less and intra-cluster similarity is more. Clustering is called as unsupervised learning because it does not use predefined classes or labels for clustering data.

Some of the requirements of clustering include scalability, ability to handle different types of data, noisy data, high dimensional data, and insensitive to the order of the input. The type of data used for the clustering algorithm includes Interval-scaled, Binary, Categorical, Ratio scaled and Attributes of mixed data types. This paper can deal with the clustering algorithm for categorical and binary data only.

There are five methods of clustering algorithms like hierarchical, partitioning, density, grid, and model-based clustering. [Fig. 1] shows the block diagram representation of the clustering methods.

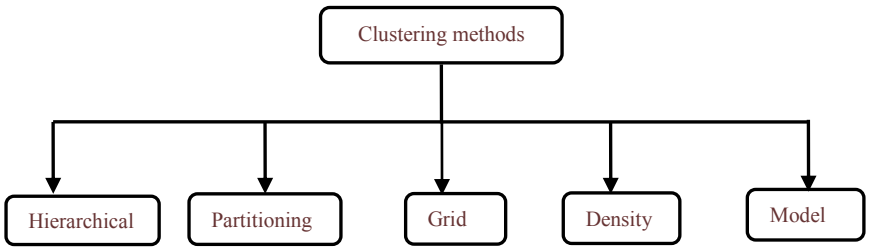


Fig. 1: Methods of clustering

The brief explanation of each of the method is discussed below

Hierarchical clustering algorithm groups the instances into a hierarchy or tree of clusters based on the distance measure as the criterion function. At the highest level, all items belong to the same cluster. Hierarchical clustering methods are of two types, the first approach is agglomerative (AGENS) or bottom-up approach. It starts with a single instance and successively combines the instances which are similar to one another, till all the clusters are merged into a single cluster or a stopping condition satisfies. The second approach is divisive (DIANA) or top-down approach. It starts with all the instances in one cluster. During every iteration, a cluster is split into smaller clusters, until all instances are in one cluster or a stopping condition satisfies [45].The tree representation of the hierarchical clustering was viewed by dendro gram.

***Corresponding Author**
Email:
valarmathi.b@vit.ac.in
Tel.: +91 9442811963

IIOAB JOURNAL
ISSN: 0976-3104

The important feature in the hierarchical clustering is that it will not assume the number of clusters. Examples of this type of clustering include ROCK, CHAMELEON, and etc.

In partitioning based clustering method, it divides a database D into 'x' partitions, where each partition corresponds to a cluster. Some examples of partitioning methods are K-means, K-medoids, and CLARANS [45].

Clusters are produced based on the density of points in density based clustering. A region with more compactness of points shows the presence of clusters, whereas regions with a low compactness of points represent the noise or outliers. DBSCAN, OPTICS, Den Clue are examples for density-based clustering [45].

In grid-based clustering, it divides the dataset consists of 'p' objects into a predetermined number of cells that form a grid structure. Few examples are STING, Wave Cluster, and CLIQUE [45].

EM, SOM, and COBWEB are the examples for model-based clustering method, in which a model is hypothesized for each of the clusters and tries to find the best fit of that model to each other in model-based clustering [45].

Some of the applications of clustering include pattern recognition, spatial data mining, World Wide Web, document clustering, image processing, cellular manufacturing, and etc.

This paper presents the literature review of the clustering algorithm for categorical and binary data, based on the type of methods used for clustering categorical data, evaluation criteria, datasets used, and input & output parameters.

[Table 1] represents the example for categorical datasets consists of nine objects, and five attributes belong to three classes. The attribute values are represented using X, Y, Z, P, Q and the corresponding attribute numbers. The whole dataset is divided into three classes namely 1, 2 and 3. The objects obj1, obj4, obj5belong to class 1, the objects obj2, obj3, obj7 belong to class 2, and similarly, the objects obj6, obj8, obj9 belong to class 3.

Table 1: A sample categorical dataset

Data Objects	$Attribute_1$	$Attribute_2$	$Attribute_3$	$Attribute_4$	$Attribute_5$	Classes
obj_1	X_1	Y_2	Z_3	P_1	Q_3	1
obj_2	X_1	Y_2	Z_3	P_1	Q_3	2
obj_3	X_2	Y_2	Z_3	P_2	Q_3	2
obj_4	X_2	Y_1	Z_1	P_2	Q_2	1
obj_5	X_1	Y_2	Z_1	P_4	Q_2	1
obj_6	X_3	Y_3	Z_1	P_4	Q_1	3
obj_7	X_2	Y_3	Z_2	P_4	Q_1	2
obj_8	X_3	Y_2	Z_2	P_3	Q_1	3
obj_9	X_3	Y_3	Z_2	P_3	Q_2	3

Steps involved in clustering

The objective of clustering is to combine the most similar objects into groups. The steps involved in the clustering are given as follows:

Step 1: Find the similarity/dissimilarity of the data objects using the distance measures
Step 2: Find the method used to form the clusters
Step 3: Decide the input parameters (e.g. Number of clusters)
Step 4: Decide the output parameters (e.g. Cluster validation measure)

Similarity and dissimilarity measures

Distance metrics are generally used to find the similarity and dissimilarity of the objects. Many benchmark distance metrics are available. Some of the commonly used distance metrics are given in [Table 2].

Table 2: Commonly used distance metrics

S.No.	Distance Measure
1.	Chebychev
2.	City block
3.	Correlation
4.	Cosine
5.	Euclidean
6.	Hamming

7.	Jaccard
8.	Mahalanobis
9.	Manhattan
10.	Minkowski
11.	Seuclidean
12.	Spearman

Other than the benchmark distance measures mentioned in [Table 1], many researchers have developed their own similarity measures. "Some of the available similarity indexes present in literature are Gower similarity (GOW) (Gower, 1971), Eskin similarity (ESK) (Eskin, Arnold, Prerau, Portmoy & Stolfo, 2002), Inverse Occurrence Frequency similarity (IOF) (Church & Gale, 1995), Occurrence Frequency similarity (OF), Good all similarity (GOO), Gambaryan similarity (GAM), Lin similarity (LIN), Anderberg similarity (AND), and Smirnov similarity (SMI). These are the similarity measures compared based on the four validation measure viz., NCC, compactness, entropy and silhouette index for 15 datasets by [35]. Whereas, [75] proposed a similarity measure MCSM (Multiple Categorical Similarity Measure) for multiple categorical datasets".

Methods for clustering data

There are five methods of clustering algorithms available in the literature "hierarchical clustering method, partition based method, density based method, grid based method, and model-based clustering method". The detailed review of the methods is given in section 3 of this paper.

Input parameters

In some clustering algorithms, the input parameter 'k' which is nothing but the number of clusters should be known before performing the clustering. Example for this type of clustering is partition based clustering methods like k-means, k-medoids, k-modes, and etc. The number of given classes are three in [Table 1]. While performing clustering the number of clusters should be given as a three. Instead, if the number of clusters is given as a two means the clustering accuracy varies automatically. So the clustering accuracy depends on the number of clusters, we should have a prior knowledge about the required number of clusters. [24] proposed an algorithm called categorical data clustering with subjective factors (CDCS). The main feature of the algorithm is automatically decides the proper clustering parameters. "[26] developed a categorical data clustering method named BK Plotto validate the clusters. (Kuo et al. 2014) developed an automatic clustering algorithm called automatic kernel clustering with bee colony optimization (AKC-BCO). It automatically decides the number of clusters and assigns data points to correct clusters".

Validation Measures

The validation measures compute the performance of the clustering algorithm. It is defined by combining compactness and separability [82]. Compactness used to measure the closeness of the cluster objects. Separability is used to measure the distinctness between the clusters. The types of validation measures available are internal validation measure and external validation measure. The first method used to evaluate the goodness of the clusters without any external information and the second method used to evaluate the clustering results by comparing the results with the externally supported class labels. The validation measures used in various studies are given in section 3.3.

Some of the commonly used internal measures are Davies-Bouldin index, Silhouette index, Bayesian information criterion (BIC), Dunn Index, etc. And the most widely used external measures are Normalized Mutual Index (NMI), Purity or Rand Index (RI), Entropy, F-measure, Adjusted Rand Index (ARI).

REVIEW METHODOLOGY

This paper presents the review based on the type of data used for clustering based on similarity or dissimilarity measure used in each article, the methods/techniques used to form clusters, the number of datasets, validation measures and tools & the system specifications of the model developed in each reference.

Methods of clustering

Cluster analysis can deal with four types of data, Interval-scaled, binary, categorical, and ratio scaled values. In this study, the focus is only on the binary and categorical data types.

Table 3: Data types incorporated in the articles

Data type	Number of articles	Articles
Binary	8	[16], [42], [43], [69], [89], [90],[96],[97]
Categorical	67	[3], [4],[6], [9],[10], [13], [11], [12], [14],[17], [18], [19],[20],[22],[21], [23], [24], [25], [26], [27],[28], [29],[31], [33], [34], [36],[37],[38], [40],[41],[44],[48],[49], [50], [51], [52],[54], [55],[56],[57], [62], [63], [64], [66], [67],[68], [70], [71], [72], [73], [74], [77],[78], [79], [80],[81], [85],[86], [87], [88],[91], [92], [93], [95], [98], [100]
Mixed numeric and categorical data	11	[1], [2], [15], [30], [32], [53], [59], [60], [61],[75], [99]

[Table 3] shows the details of the articles which have used the pure categorical data, pure binary data, and mixed numeric & categorical data types. The detailed review of the clustering algorithms with respect to the basic methods or division of clustering is discussed in this section.

Partition based method

Many algorithms were available in the literature for clustering larger datasets. The algorithms like "CLARANS proposed by Ng and Han (1994), BRICH by [100], and DBSCAN by (Ester et al. 1996) are suitable for solving numerical datasets" only and not applicable for solving categorical dataset [55].

Ralambondrainy (1995) proposed a categorical clustering algorithm using k-means algorithm by converting the categorical values into binary values. This approach treats the binary values as numeric values and performs "k-means clustering". The disadvantage of this algorithm includes the "computational cost" and the mean values between 0 and 1 do not signify the uniqueness of the clusters.

[55] proposed two algorithms for clustering categorical data by extending the k-means algorithm. First one is k-modes algorithm by replacing mean by mode, it used a simple matching distance measure for clustering categorical attributes. In order to reduce the computational cost a "frequency based method" was used to recalculate the modes. Second is "k-prototype algorithm" by combining "k-means and k-modes algorithm" for clustering data with mixed numeric and categorical attributes.

Many partitioning based clustering algorithms required a random selection or pre-setting of initial points (mean or modes) of the clusters for clustering. Choosing of these initial points randomly will leads to different cluster results. So, [88] did an experimental study on the refinement of initial points to "k-modes" type categorical clustering algorithm for the better clustering results. Based on the experimental study they found that, k-populations algorithm produced better clustering results.

[75] mentioned that the performance of k-modes, k-prototypes and fuzzy k-modes algorithms results in local optimum only. So, a tabu search method for obtaining the global optimum results for categorical data is proposed.

[63] developed a new fuzzy based clustering method by extending "fuzzy k-modes" algorithm for clustering categorical data. In that, the hard-type centroids were replaced by fuzzy centroids in order to fully exploit the power of fuzzy sets. The proposed method was compared with the two existing algorithms namely "k-modes and fuzzy k-modes" and reported that it produced better clustering results.

[2] developed a "k-means" type model for categorical and numeric data clustering. The modified description of the initial points was introduced to conquer the numeric data alone constraint of the traditional "k-means" algorithm. A novel cost function and a dissimilarity measure were also proposed. The proposed algorithm was tested on real life datasets.

"[9] developed a clustering algorithm for handling high-dimensional categorical data by extending the "k-modes" algorithm using optimization methods". [9, 10, 11] experimented a k-mode type algorithm which automatically initializes the cluster centers and the number of clusters. Similarly, (He et al.2011) useda "k-modes algorithm using attribute value weighting" in the distance computation.

(Hatamlou 2012) introduced a new partitioning based algorithm using the concept of binary search algorithm. The initial centroids were chosen from the different parts of the dataset. It is noted that it converged to the same results in different runs. [15] proposed a geometric codification for clustering mixed categorical and numeric data. It codified the categorical attributes into numerical values and performed numerical clustering algorithm by combined with k-means algorithm.

[21] developed a new distance measure and a rough membership function to overcome the limitation of simple matching distance measure and Ng's distance measure for the k-modes algorithm for clustering categorical data. "[12] proposed a "weighting k-modes algorithm" for categorical data to perform subspace

clustering". In addition to the usual k-modes clustering procedure, a step to calculate weights automatically using complement entropy for all the dimensions in each and every cluster was added.

[60] developed another version of "k-prototype algorithm" for clustering "numeric and categorical data". To represent a prototype of clusters the mean & fuzzy centroids were combined and in order to calculate the distance among instances and the prototypes a distance measure was developed. "Similarly, (Ji et al. 2013) developed an improved k-prototype algorithm for mixed numeric and categorical data", here the prototype of the "categorical attributes in the cluster was represented by distribution centroids and to represent the prototype of a numerical attributes in a cluster the mean and the distribution centroids". A new dissimilarity measure was proposed to find the distance between the instances and the prototypes. In both methods, the performance of the algorithm was tested for four real world datasets and the results were compared with the traditional clustering algorithms.

"[86] proposed a medoids based clustering method called k-Approximate Modal Haplotype (k-AMH). k-AMH is a medoids based clustering for clustering categorical data and it was compared with the centroids based clustering methods like k-modes, k-population, and fuzzy k-modes algorithm in terms of clustering accuracy. [87] enhanced the k-AMH algorithm using the same procedure as that of k-AMH, termed as (Nk-AMH I), (Nk-AMH II),and (Nk-AMH III) but with the addition of two methods likely new initial center selection and new dominant weighting methods for clustering categorical data based on optimization and fuzzy procedures.

[11] proposed a fuzzy clustering algorithm by modifying the objective function of the fuzzy k-modes algorithm by including between cluster information to minimize the within cluster dispersion and between cluster partition simultaneously. [7] proposed a k-modes type clustering algorithm for categorical data. The objective function is modified by adding the between cluster similarity term in it, to overcome the limitation of weak separation of clusters in usual clustering algorithms. The algorithm was tested for some real world datasets and reported that this method produced better results than original counterparts in categorical data clustering and applicable for large datasets.

[92] compared the performance of the objective functions of the algorithms like k-medoids, k-modes, and within cluster dispersion analytically. Also, they verified the objectives for real valued datasets. The experiments were conducted to prove the performance of the objective function using the real-life data sets and reported that within cluster dispersion algorithm performs better than other methods two methods. Similarly, [8] compared the generalization, effectiveness and normalization objective functions of the internal validity functions like k-modes, category utility function, and the information entropy function by using the developed generalized validity function for evaluating the categorical data results in a solution space. Also, they addressed the problem while using these validity functions for evaluating the clusters whether the between cluster information is ignored".

Hierarchical based methods

"ROCK, a robust clustering algorithm for categorical and binary data using links was proposed by (Guha et al.2000)". It overcomes the drawbacks of the traditional clustering algorithms using distance measure or similarity measure. Using distance measure or similarity measure for clustering categorical and binary data is not appropriate. So, the concept of 'link' was introduced to find the common neighbors between the data points. The performance of the algorithm was tested on three datasets like, mushroom, congressional votes, US Mutual funds. "A quick version of the ROCK algorithm called QROCK" was proposed by [36] based on the concept of graphs. The final clusters were the components of the graph and the data points as the vertices. The main advantage of the QROCK over the ROCK algorithm was, the computation time of QROCK was reduced because of the 'merge' and 'find' concept introduced in the ROCK algorithm.

"[89] proposed a hierarchical clustering algorithm for binary gene expression data". [5] developed a scalable clustering algorithm called LIMBO, a bottleneck information framework for the design of the novel distance measure for the categorical attributes was used. It is a kind of hierarchical clustering algorithm. The main advantage of the LIMBO was in single execution and it could produce the clusterings of different sizes.

(Barbara et al. 2002) proposed COOLCAT, clustering algorithms for the categorical data based on entropy. It is applicable for both categorical data and also data streams. Entropy is lower for clusters having similar objects and it is higher for clusters having dissimilar objects.

"[57] proposed a framework to learn a context-based dissimilarity measure for categorical attributes". Based on the distribution of objects in other attributes, the distance between two objects of an attribute is determined. This method is embedded in hierarchical clustering method to validate the proposed method.

[93] developed a divisive hierarchical clustering termed as DHCC. The task of categorical data clustering was viewed in the type of optimization point of view and proposed a procedure for initialization and splitting of clusters. The advantages of this method is, it performs automatic clustering, "the dendro gram representation is obtained due to the hierarchical nature of the algorithm, the order of the data is independent, scalable for large dataset, and finding clusters in subspaces".

"[80] proposed an information theory based hierarchical divisive clustering algorithm for categorical data using the mean gain ratio (MGR) of the attributes. The attribute having highest MGR is selected as the clustering attributes and equivalence class with minimum entropy is determined as the cluster and the other equivalence class is considered as the new dataset and repeats the process until all the instances are grouped into the clusters. The performance of the MGR was compared with the existing other four algorithms based on the entropy or mutual information such as COOLCAT (Barbara et al. 2002), MMR [79], K-ANMI [51], G-ANMI [33]".

Density-based clustering methods

[4] enters proposed a hierarchical density-based clustering method for categorical data named as HIERDENC and also developed a complementary index for searching dense subspaces. In that, the data was represented in the form of cube, where there is no ordering of the instances. Because of this advantage if the new instance enters into the system the HIERDENC index is only updated and the re clustering was not required. Initially the formation of clusters was started from the dense regions of the cube. Later the close by dense regions was connected to form further clusters. The HIERDENC method was compared with few existing categorical clustering algorithms and reported that the algorithm performed better scalability, runtime and cluster quality on large datasets.

[13] proposed an enhanced DBSCAN algorithm for incrementally building and updating of arbitrarily shaped clusters in large datasets. Instead of searching the whole dataset it searches only the partitions, this leads to the betterment of the results when compared with the other incremental clustering algorithms.

Model-based methods

One-dimensional Clustering is nothing but clustering data by considering all the attributes in the dataset. This way of clustering is not appropriate for complex datasets with many attributes. To overcome this draw back "[29] proposed a model based method for clustering multidimensional categorical data".

(Bauldry et al. 2015) proposed a model based algorithm which it directly finds the number of clusters and also can handle the external variables. [91] proposed a model based method based on the mixture of latent trait models with common slope parameters for clustering binary data. To determine the model parameters by means of fast algorithms the various approximations to the likelihood is exploited.

Artificial intelligence based methods

"The fuzzy k-modes algorithm is efficient for clustering categorical data. The fuzzy objective function is minimized when the algorithm searches for the fuzzy membership matrix. So, the fuzzy k-modes algorithm may stop at local optimal solution. To overcome the drawback of the fuzzy k-modes algorithm [38] proposed a genetic fuzzy k-modes algorithm for clustering categorical data. Where, the GA and fuzzy k-modes algorithms were hybridized to find the global optimal solutions. This algorithm was tested for two real life datasets and the performance was compared".

Many researches were found solutions for the categorical data clustering using the single measure for finding the clusters. This may not suitable for different datasets. To overcome this [74] developed a multi-objective genetic algorithm based fuzzy clustering for categorical data. "The two objective functions optimized by the proposed method are fuzzy compactness and the fuzzy separation of the clusters". This method was compared qualitatively and quantitatively with other algorithms and also it was tested for synthetic and real life datasets.

[66] proposed a self-organization map (SOM) for clustering and visualization of categorical data based on the Kohonen map. [53] proposed an extended SOM called MixSOM algorithm for clustering mixed numeric and categorical data.

Few authors proposed a single objective function for clustering categorical data. Such single objective function may be inappropriate for all type of datasets. So, in order to overcome this drawback [84] developed a multi objective incremental learning evolution based fuzzy clustering algorithm for clustering categorical data. The evolution based fuzzy clustering method was combined with random forest classifier for categorical clustering. This algorithm was tested for "synthetic and real world datasets to show the performance of the algorithm".

In many SOM, categorical data cannot be directly processed. It should be converted into a binary value before processing. [32] developed a SOM architecture which processes categorical data without any conversion to binary values.

[85] developed a clustering algorithm by combining "rough set and fuzzy set theories". "An ensemble based framework is designed to find the best clustering results for different categorical data sets".

Other approaches

IIOAB JOURNAL
ISSN: 0976-3104

[43] used Bernoulli distribution mixtures for the cluster analysis with binary data, and the results were compared with the Monte-Carlo numerical experiments. [90] proposed an extension of Latent class analysis model for improving the clustering accuracy in each cluster and used Bernoulli distribution mixtures to solve the difficulties of the clustering problem, i.e. to find the number of clusters and to find the correlation matrix for each cluster, etc.
CACTUS proposed by [39], is a summarization based clustering method for categorical dataset with large number of attributes. It required only two scan of the dataset for the formation of clusters and it performs subsace clustering to find the clusters in the subset of attributes. It is a three phase algorithm, first is a summarization phase, second one is a clustering phase and the last phase is a validation phase. The performance of CACTUS was tested for real life and synthetic datasets and it was compared with the existing algorithms.

Squeezer a clustering algorithm proposed by [98] for categorical attributes is suitable for clustering data streams. This algorithm is suitable for solving small dataset only. For handling of large datasets, they proposed an enhanced algorithm called d-squeezer. SCLOPE is also a clustering algorithm for categorical data streams proposed by (Ong et al. 2003).

(He et al. 2005) considered the commonalities between the two different research problems, categorical data clustering and the cluster ensembles. They developed an algorithm based on cross-fertilization between a two problems for clustering categorical data. Whereas, [56] proposed a link based approach for solving the above said two problems.

"[79] proposed an algorithm for clustering categorical data based on the rough set theory called min-min roughness (MMR). The MMR can handle the uncertainty in the clustering process.

[68] proposed a hierarchical clustering algorithm for categorical data based on the rough set model. ATMDP (Total Mean Distribution Precision) method for selecting the partitioning attribute based on probabilistic rough set theory also developed. Based on the TMDP a clustering algorithm called MT MDP (Maximum Total Mean Distribution Precision) was developed. The performance of the MT MDP was compared with the MMR algorithm and claimed that MT MDP algorithm was superior to the MMR algorithm.

[73] compared with some of the existing categorical clustering algorithms using Monte Carlo simulation. The algorithms are like average linkage, ROCK, k-modes, fuzzy k-modes and k-populations were compared.

[67] developed a dissimilarity measure termed as CATCH (an effective Categorical data dissimilarity measure using a distributional Characteristic in High-dimensional space) for clustering categorical data. Zhang and Gu (2014) developed a similarity measure and a affinity propagation (AP) algorithm for clustering mixed data types.

[95] developed a k-modes type clustering algorithm for categorical data which improves the quality of the clusters by using non-dominated sorting genetic algorithm-fuzzy membership chromosome (NSGA-FMC) which combines fuzzy genetic algorithm and multi-objective optimization. Park and Choi (2015) proposed a roughest based approach for clustering categorical dataset named information-theoretic dependency roughness (ITDR).

[26] proposed a method called Maximal Resemblance Data Labeling (MARDL) for clustering concept drifting categorical data. For the concept drifting an algorithm named DCD Detecting concept, drift was also developed. The objective of the algorithm was to find the difference between the distributions of the clusters of the current clustering subset and the last subset. It decides whether the re-clustering was required or not. (Reddy etal.2014) developed a method for data labeling and the concept drift detection based on the entropy model in rough set theory.(Li Y et al. 2014) proposed a three dissimilarity measures based on incremental entropy and an integrated framework consists of a three algorithms for clustering categorical data streams with concept drift.

Many subspace clustering algorithms were proposed for clustering categorical datasets. Subspace clustering is used to find clusters within the datasets in different subspaces (Parson et al. 2004). [3, 1,12, 40, 29] developed a subspace clustering algorithm for categorical data.

(Hatamlou 2013) developed an optimization algorithm named Black hole for data clustering. Black hole algorithm also starts with the initial population solutions for an optimization problem like other population-based methods. In all iterations the best candidate was selected to the black hole. (Hautamäki et al. 2014) proposed a novel clustering algorithm based on alocal search for the objective function. The expected entropy was considered as the objective function for this algorithm. The results were compared with the existing six iterative clustering algorithms and showed that the proposed method produced the best clustering results than the other six methods".

Comparison of various clustering methods

[Table 4] describes the comparison of various methods with respect to the following criteria.

IIOAB JOURNAL
ISSN: 0976-3104

- K : The number of clusters known apriori.
  - The value in the table is YES if the cluster number is known at the beginning of the algorithm else, the value is NO
- N: Number of datasets solved
- LD: Largest size of the dataset solved
- S: Whether synthetic datasets generated and tested
  - The value in the table is YES if a dataset is generated and tested else, the value is NO
- C: Compared with the existing methods
  - The value in the table is YES if it is compared with the existing methods otherwise, the value is NO
- Software's or programming languages used for implementing the algorithm in various research articles.

From the literature, it is clear that most of the algorithms required the number of clusters as input, very few algorithms only automatically decides the number of clusters. In the same way, many partitioning based clustering algorithms required a random selection or pre-setting of initial points (mean or modes) of the clusters for clustering. Choosing of these initial points randomly will leads to different cluster results. So, the k-populations algorithm emerged to "automatically initialize the cluster centers and the number of clusters, which leads to the better clustering results [88]".Each clustering algorithm has its own merits and demerits. There is no common clustering algorithm available for handling all kinds of data types. One dimensional clustering by considering all the attributes in the dataset for clustering categorical data is not appropriate for complex datasets so the multi-dimensional clustering was proposed by [29].

Table 4: Comparison of various clustering methods

S. No	Source	K	N	LD	S	C	Impl. Tools
1.	Zhang T et al.(1996)	NA	NA	NA	NO	NO	-
2.	Huang Z (1998)	YES	2	690	YES	YES	-
3.	Ganti V et al. (1999)	YES	2	30919	NO	YES	-
4.	Karypis G, Han ES (1999)	YES	5	10000	NO	YES	-
5.	Guha S et al. (2000)	YES	3	8124	YES	YES	-
6.	Barbará D et al. (2002)	YES	3	1000	YES	YES	-
7.	Ng MK, Wong JC (2002)	YES	4	690	NO	YES	C++
8.	Sun Y et al. (2002)	YES	1	47	NO	YES	-
9.	Zengyou H et al. (2002)	NO	2	8124	YES	YES	Java
10.	Szeto LK et al. (2003)	NO	1	6178	NO	YES	-
11.	Andritsos P et al. (2004)	YES	3	8124	YES	YES	-
12.	Kim DW, et al. (2004)	YES	3	202	NO	YES	-
13.	Ong K et al. (2004)	YES	4	990,002	YES	YES	C
14.	Chang CH, Ding ZK (2005)	YES	5	8124	NO	YES	-
15.	Dutta M et al.((2005)	YES	2	8124	NO	YES	-
16.	He Z, Xu X, Deng S (2005)	YES	4	8124	NO	YES	
17.	Kim DW et al.(2005)	YES	4	202	NO	YES	-
18.	Li T (2005)	YES	6	8280	NO	NO	-
19.	Ahmad A, Dey L (2007)	YES	4	690	NO	YES	-
20.	Cesario E et al. (2007)	YES	13	8124	YES	YES	C++
21.	Parmar D et al. (2007)	YES	3	8124	NO	YES	VB.Net
22.	He Z et al.((2008)	YES	3	8124	NO	YES	-
23.	Andreopoulos B et al.((2009)	NO	5	12960	NO	YES	Python

"K- Number of clusters known Apriori; N- Number of real life dataset solved; LD-Largest size of the dataset; S- Synthetic datasets used;

C- Compared with existing algorithms; Impl. Tools- Implementation tools; NA- Not Available"

Table 4: Comparison of various clustering methods (continued)

S. No	Source	K	N	LD	S	C	Impl. Tools
1.	Cao F et al. (2009)	YES	4	8124	NO	YES	-
2.	Chen HL et al.(2009)	NA	NA	493,857	NO	NA	-
3.	Chen K, Liu L (2009)	NO	1	2,458,284	YES	YES	-
4.	Gan G et al. (2009)	YES	2	435	NO	YES	C++
5.	Mukhopadhyay, A et al.((2009)	YES	4	699	YES	YES	MATLAB
6.	Aranganayagi S, Thangavel K (2010)	NO	4	8124	NO	YES	-
7.	Deng S et al.(2010)	YES	4	8124	NO	YES	Java
8.	Tamhane AC, Qiu D, Ankenman BE (2010)	YES	2	10658	YES	YES	C++
9.	Ahmad A, Dey L (2011)	YES	4	8124	NO	YES	-
10.	Bai L et al. (2011)	YES	4	8124	NO	YES	-
11.	Bai L et al. (2011)	YES	7	2,458,284	YES	YES	-
12.	Cao F, Liang J (2011)	YES	1	8124	NO	YES	-

13.	He Z et al. (2011)	YES	5	12690	NO	YES	Java
14.	Rendón E et al.(2011)	YES	NA	NA	YES	YES	-
15.	Bai L et al.(2012)	YES	6	67,557	NO	YES	-
16.	Barcelo-Rico F, Diez JL (2012)	YES	6	30161	NO	YES	-
17.	Cao F et al. (2012)	YES	5	12690	NO	YES	MATLAB
18.	Chen T et al. (2012)	YES	32	20000	YES	YES	Java
19.	Hatamlou A (2012)	YES	6	1473	NO	YES	-
20.	Hsu CC, Lin SH (2012)	YES	2	48842	YES	YES	-
21.	Iam-On N et al. (2012)	YES	9	100000	NO	YES	-

"K- Number of clusters known Apriori; N- Number of real life dataset solved; LD-Largest size of the dataset; S- Synthetic datasets used;
C- Compared with existing algorithms; Impl. Tools- Implementation tools; NA- Not Available"

Table 4: Comparison of various clustering methods (continued)

S. No	Source	K	N	LD	S	C	Impl. Tools
1.	Reddy HV et al. (2014)	YES	NA	NA	YES	NO	-
2.	Saha I, Maulik U (2014)	YES	4	690	YES	YES	-
3.	Zhang K, Gu X (2014)	NO	4	690	NO	YES	C
4.	Bai L, Liang J (2015)	YES	12	8124	NO	YES	-
5.	Bakr AM et al. (2015)	YES	6	20000	NO	YES	-
6.	Baudry JP et al. (2015)	YES	3	440	NO	NA	-
7.	Bouguessa M (2015)	YES	3	8124	YES	YES	-
8.	Del Coso C et al. (2015)	YES	5	48842	NO	YES	-
9.	Dos Santos TRL, Zárate LE (2015)	NA	15	893	NO	YES	-
10.	García-Magariños M, Vilar J (2015)	YES	1	6000	YES	YES	R
11.	Park IK, Choi GS (2015)	YES	1	101	NO	YES	MATLAB
12.	Saha I et al. (2015)	YES	4	435	YES	YES	MATLAB
13.	Seman A et al. (2015)	YES	5	699	NO	YES	-
14.	Tang Y et al. (2015)	YES	2	2400	NO	YES	-
15.	Yang CL et al. (2015)	YES	3	435	NO	YES	MATLAB
16.	Chen LiFei et al. (2016)	YES	4	3190	YES	YES	-

"K- Number of clusters known Apriori; N- Number of real life dataset solved; LD-Largest size of the dataset; S- Synthetic datasets used;
C- Compared with existing algorithms; Impl. Tools- Implementation tools; NA- Not Available"

Frequently used datasets

The real life data set repositories available for clustering are "Frequent Item set Mining Dataset Repository (FIMI), University of California Irvine Machine Learning Repository(UCI)", their URL's are given as follows

"FIMI - (http://fimi.cs.helsinki.fi/testdata.html)
UCI – (http://www.ics.uci.edu/mlearn/MLRepository.html)".

[Table 5] shows a ten frequently used real life datasets with the number of objects and the number of attributes.

Table 5: Frequently used real life datasets

S. No.	Datasets	No. of instances	No. of attributes
1	Soybean	47	35
2	Zoo	101	16
3	Heart Disease	303	13
4	Dermatology	366	33
5	Congressional votes	435	16
6	Credit Approval	690	15
7	Wisconsin Breast Cancer	699	9
8	Car evaluation	1728	4
9	Chess	3196	36
10	Mushroom	8124	22

Validation Measures

[82] compared six internal indexes such as Bayesian information criterion (BIC), Calinski-Harabasz (CH), Davies - Bouldin (DB),Silhouette(SIL), Novel Validity index(NIVA) and DUNN index and four external indexes such as purity, Entropy, F-measure, and Normalized mutual index (NMI) for 13 datasets. The clusters for the comparison were obtained by the k-means and Bisecting-K means algorithms and reported that the internal

IIOAB JOURNAL ISSN: 0976-3104

measures are more accurate than the external measures. [Table 6] and [Table 7] show the external and the internal validation measures used in the evaluation of the clustering in different studies respectively.

Table 6: External validation measures incorporated in various studies

S.No	External Validation measure	Articles
1.	Clustering Accuracy / Purity	[3],[6],[7],[9],[11],[12],[15],[18],[19],[20],[21], [22], [23], [24],[30],[32],[37],[41], [44], [52],[53], [55], [56],[57],[60],[61],[63], [64],[65], [68],[69],[70],[75],[76],[77], [79],[81],[82],[83],[87],[88]
2.	Adjusted Rand Index	[10],[18],[21], [41],[57],[68],[74],[83],[91],[95]
3.	Number of correctly classified instances	[71]
4.	Micro-right	[71]
5.	Confusion matrix	[49],[67]
6.	Normalized Mutual Information	[29],[56], [57],[68],[82]
7.	Precision or Recall or F-measure or micro precision	[1], [4],[7],[9],[11],[12],[20],[25],[31],[32],[46],[49],[82],[87],[99]
8.	Error rate	[10],[23], [37], [38], [47], [55],[65],[98]
9.	Average clustering error	[2], [33],[50], [51]
10.	Gain ratio	[53]
11.	Category utility	[5], [8], [10],[14], [25]
12.	CPU time	[4], [36],[93], [95]
13.	Jaccard	[18]
14.	Fowlkes	[18]
15.	Entropy	[14],[35],[48],[53],[82]

Table 7: Internal validation measures incorporated in various studies

S. No	Internal Validation measure	Articles
1.	Dunn index	[82],[84]
2.	Silhouette	[35],[41],[82]
3.	Davies-Bouldin index(DB)	[82],[84]
4.	Bayesian information criterion(BIC)	[82],[91]
5.	Novel Validity Index(NIVA)	[82]
6.	Calinski-Harabasz index	[82]
7.	Percentage of correct pair (%CP)	[83], [84]
8.	Minkowski score (MS)	[83],[84]
9.	Compactness	[35], [95]
10.	Gavrilov index (GI)	[41]

Tools for performing clustering

There are few software's available for performing some data mining techniques including clustering. Some of the software's are open source, and few are proprietary version. The details of the commonly used software's for clustering are given in [Table 8].

Table 8: Most widely used open source software's

S. No	Software	Type	Year
1.	CLUTO	Open Source	2002
2.	gCLUTO	Open Source	2003
3.	MALLET	Open Source	2011
4.	Mlpy	Open Source	2012
5.	Orange	Open Source	2009
6.	R-'cluster' Package- CRAN, Rattle	Open Source	2011
7.	TANAGRA	Open Source	2004
8.	wCLUTO	Open Source	2003
9.	Waikato Environment for Knowledge Analysis (Weka)	Open Source	1993
10.	MATLAB- Clustering tool box	Proprietary	1984
11.	Origin	Proprietary	1993
12.	RapidMiner	Proprietary	2001
13.	Statistical Analysis System(SAS)	Proprietary	1971
14.	SPSS	Proprietary	1968

IIOAB JOURNAL
ISSN: 0976-3104

CONCLUSION

This paper provides the overview of the methods used for clustering categorical clustering data like similarity or dissimilarity measures, validation measures available in the literature, and available real life categorical datasets experimented in the different studies. Most of the authors incorporated partition based and hierarchical based methods for clustering categorical data. Partition based clustering is suitable for all types of data, the only drawback of this method is, the number of clusters must be known apriori. This may overcome by choosing the random number of clusters and then increase or decrease the number of clusters to certain level based on the accuracy. Subspace clustering method is appropriate for clustering high-dimensional categorical data. There are very few algorithms only available for model-based and density-based clustering methods and also few papers only incorporated the optimization techniques for categorical data clustering. A small number of heuristics or meta heuristics methods only available for clustering categorical data. The scope for the future research includes the formation of algorithms based on evolutionary algorithms like Genetic Algorithm, Simulated Annealing, and etc. for clustering categorical data.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Ahmad A, Dey L. [2011] A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. Pattern Recognition Letters. 32:1062–1069.
[2] Ahmad A, Dey L. [2007] A k-mean clustering algorithm for mixed numeric and categorical data. Data and Knowledge Engineering. 63:503–527.
[3] Al-Razgan M, Domeniconi C, Barba D. [2008] Random subspace ensembles for clustering categorical data. Supervised and Unsupervised Ensemble Methods and their Applications. Studies in Computational Intelligence. 126:31- 48.
[4] Andreopoulos B, An A, Wang X, Labudde D. [2009] Efficient layered density-based clustering of categorical data. Journal of Biomedical Informatics. 42:365–376.
[5] Andritsos P, Tsaparas P, Miller RJ, Sevcik KC. [2004] LIMBO: Scalable clustering of categorical data. Advances in Database Technology-EDBT. 2992:123–146.
[6] Aranganayagi S, Thangavel K. [2010] Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure. International Journal of Information and Mathematical Sciences. 6: 21-29.
[7] Bai L, Liang J. [2014] The k-modes type clustering plus between-cluster information for categorical data Neurocomputing. 133:111–121.
[8] Bai L, Liang J. [2015] Cluster validity functions for categorical data: a solution-space perspective. Data Mining and Knowledge Discovery. 29:1560–1597.
[9] Bai L, Liang J, Dang C. [2011] an initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. Knowledge-Based Systems. 24:785–795.
[10] Bai L, Liang J, Dang C, Cao F. [2011] A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognition. 44:2843–2861.
[11] Bai L, Liang J, Dang C, Cao F. [2012] A cluster centers initialization method for clustering categorical data. Expert Systems with Applications. 39:8022–8029.
[12] Bai L, Liang J, Dang C, Cao F. [2013] A novel fuzzy clustering algorithm with between-cluster information for categorical data. Fuzzy Sets and Systems. 215:55–73.
[13] Bakr AM, Ghanem NM, Ismail MA. [2015] efficient incremental density-based algorithm for clustering large datasets. Alexandria Engineering Journal. 54:1147–1154.
[14] Barbará D, Couto J, Li Y. [2002] COOLCAT: An entropy-based algorithm for categorical clustering In: Proceedings of11th ACM international conference on information and knowledge management, McLean, VA, USA. 582-589.
[15] Barcelo-Rico F, Diez JL. [2012] Geometrical codification for clustering mixed categorical and numerical databases. Journal of Intelligent Information Systems. 39:167–185.
[16] Barthelemy J, Brucker F. [2008] Binary clustering. Discrete Applied Mathematics. 156:1237–1250.
[17] Baudry JP, Cardoso M, Celeux G, Amorim MJ, Ferreira AS. [2015] Enhancing the selection of a model-based clustering with external categorical variables. Advances in Data Analysis and Classification. 1:1–20.
[18] Bouguessa M. [2015] Clustering categorical data in projected spaces. Data Mining and Knowledge Discovery. 29:3–38.
[19] Cao F, Liang J. [2011] A data labeling method for clustering categorical data. Expert Systems with Applications. 38:2381–2385.
[20] Cao F, Liang J, Bai L. [2009] A new initialization method for categorical data clustering. Expert Systems with Applications. 36:10223–10228.
[21] Cao F, Liang J, Li D, Zhao X. [2013] A weighting k-modes algorithm for subspace clustering of categorical data. Neurocomputing. 108:23–30.
[22] Cao F, Liang J, Li D, Bai L, Dang C. [2012] A dissimilarity measure for the k-Modes clustering algorithm. Knowledge-Based Systems. 26:120–127.
[23] Cesario E, Manco G, Ortale R. [2007] Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data. IEEE Transactions on Knowledge and Data Engineering. 19:1607–1624.
[24] Chang CH, Ding ZK. [2005] Categorical data visualization and clustering using subjective factors. Data and Knowledge Engineering. 53:243–262.
[25] Chen HL, Chen MS, Lin SC. [2009] Catching the trend: A framework for clustering concept-drifting categorical data. IEEE Transactions on Knowledge and Data Engineering. 21:652–665.
[26] Chen K, Liu L. [2009] "Best K": critical clustering structures in categorical datasets. Knowledge and Information Systems. 20:1–33.
[27] Chen LiFei. [2015] A probabilistic framework for optimizing projected clusters with categorical attributes. Science China information sciences. 58:1-15.
[28] Chen LiFei, Wang S, Wang K, Zhu J. [2016] Soft subspace clustering of categorical data with probabilistic distance. Pattern Recognition. 51:322–332.
[29] Chen T, Zhang NL, Liu T, Poon KM, Wang Y. [2012] Model-based multidimensional clustering of categorical data. Artificial Intelligence. 176:2246–2269.

[30] Cheung YM, Jia H. [2013] Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognition. 46:2228–2238.
[31] Çilingirtürk AM, Ergüt Ö. [2014] Hierarchical Clustering with Simple Matching and Joint Entropy Dissimilarity Measure. Journal of Modern Applied Statistical Methods. 13:329-338.
[32] Del Coso C, Fustes D, Dafonte C, Nóvoa FJ, Rodríguez-Pedreira JM, Arcay B. [2015] Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons. Applied Soft Computing. 36:246–254.
[33] Deng S, He Z, Xu X. [2010] G-ANMI: A mutual information based genetic clustering algorithm for categorical data. Knowledge-Based Systems. 23:144–149.
[34] Do H J, Kim JY. [2009] Clustering categorical data based on combinations of attribute values. International Journal of Innovative Computing. Information and Control. 5:4393–4405.
[35] Dos Santos TRL, Zárate LE. [2015] Categorical data clustering: What similarity measure to recommend. Expert Systems with Applications. 42:1247–1260.
[36] Dutta M, Mahanta AK, Pujari AK. [2005] QROCK: a quick version of the ROCK algorithm for clustering of categorical data. Pattern Recognition Letters. 26:2364–2373.
[37] Elavarasi SA, Akilandeswari J. [2014] Occurrence based categorical data Clustering using cosine and binary matching Similarity measure. Journal of Theoretical and Applied Information Technology. 68:209-214.
[38] Gan G, Wu J, Yang Z. [2009] A genetic fuzzy k-Modes algorithm for clustering categorical data. Expert Systems with Applications. 36:1615–1620.
[39] Ganti V, Gehrke J, Ramakrishnan R. [1999] CACTUS–Clustering Categorical Data Using Summaries. In: Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, CA, USA. 73–83.
[40] Gao C, Pedrycz W, Miao D. [2013] Rough subspace-based clustering ensemble for categorical data. Soft Computing. 17:1643–1658.
[41] García-Magariños M, Vilar J. [2015] A framework for dissimilarity-based partitioning clustering of categorical time series. Data Mining and Knowledge Discovery. 29:466–502.
[42] Gebhardt F. [1999] Cluster tests for geographical areas with binary data. Computational Statistics and Data Analysis. 31:39–58.
[43] Govaert G, Nadif M. [1996] Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. Computational Statistics & Data Analysis. 23:65–81.
[44] Guha S, Rastogi R, Shim K. [2000] ROCK: A Robust Clustering Algorithm for Categorical Attributes. Information Systems. 25:345–366.
[45] Han J, Kamber M, Pie J. [2011] Data Mining concepts and techniques, 3rd edition, Morgan Kaufmann publishers Inc. San Francisco, CA, USA.
[46] Hatamlou A. [2012] In search of optimal centroids on data clustering using a binary search algorithm. Pattern Recognition Letters. 33:1756–1760.
[47] Hatamlou A. [2013] Black hole: A new heuristic optimization approach for data clustering. Information Sciences. 222:175–184.
[48] Hautamäki V, Pöllänen A, Kinnunen T, Lee KA, Li H, Fränti P. [2014] A Comparison of Categorical Attribute Data Clustering Methods.LNCS. 8621:53–62.
[49] He X, Feng J, Konte B, Mai ST, Plant C. [2014] Relevant Overlapping Subspace Clusters on Categorical Data Categories and Subject Descriptors. In: Proceedings of the 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD). 213–222.
[50] He Z, Xu X, Deng S. [2005] A cluster ensemble method for clustering categorical data. Information Fusion. 6:143–151.
[51] He Z, Xu X, Deng S. [2008] k-ANMI: A mutual information based clustering algorithm for categorical data. Information Fusion. 9:223–233.
[52] He Z, Xu X, Deng S. [2011] Attribute value weighting in k-modes clustering. Expert Systems with Applications. 38:15365–15369.
[53] Hsu CC, Lin SH. [2012] Visualized analysis of mixed numeric and categorical data via extended self-organizing map. IEEE Transactions on Neural Networks and Learning Systems. 23:72–86.
[54] Huang W, Pan Y, Wu J. [2012] Goodman-Kruskal measure associated clustering for categorical data. International Journal of Data Mining, Modeling and Management. 4: 334-360.
[55] Huang Z. [1998] Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 2:283–304.
[56] Iam-On N, Boongeon T, Garrett S, Price C. [2012], A link-based cluster ensemble approach for categorical data clustering. IEEE Transactions on Knowledge and Data Engineering. 24:413–425.
[57] Ienco D, Pensa RG, Meo R. [2012] From Context to Distance. ACM Transactions on Knowledge Discovery from Data. 6:1–25.
[58] Jain A, Murty M, Flynn F. [1999] Data Clustering: A Review, ACM Computing Surveys. 31:264-323.
[59] Ji J, Pang W, Zheng Y, Wang Z, Ma Z. [2015] An Initialization Method for Clustering Mixed Numeric and Categorical Data Based on the Density and Distance. International Journal of Pattern Recognition and Artificial Intelligence. 29:1550024.
[60] Ji J, Pang W, Zhou C, Han X, Wang Z. [2012] A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowledge-Based Systems. 30:129–135.
[61] Ji J, Bai T, Zhou C, Ma C, Wang Z. [2013] An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neuro computing. 120:590–596.
[62] Karypis G, Han ES. [1999] CHAMELEON : A Hierarchical Clustering Algorithm Using Dynamic Modeling. IEEE Computer. 32:68–75.
[63] Kim DW, Lee KH, Lee D. [2004] Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters. 25:1263–1271.
[64] Kim DW, Lee K, Lee D, Lee KH. [2005] A k-populations algorithm for clustering categorical data. Pattern Recognition. 38:1131–1134.
[65] Kuo RJ, Huang YD, Lin CC, Wu YH, Zulvia FE. [2014] Automatic kernel clustering with bee colony optimization algorithm. Information Sciences. 283:107–122.
[66] Lebbah M, Benabdeslem K. [2010] Visualization and clustering of categorical data with probabilistic self-organizing map. Neural computing and applications. 19:393-404.
[67] Lee J, Lee YJ. [2014] An effective dissimilarity measure for clustering of high-dimensional categorical data. Knowledge and Information Systems. 38:743–757.
[68] Li M, Deng S, Wang L, Feng S, Fan J. [2014] Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. Knowledge-Based Systems. 65:60–71.
[69] Li T. [2005] A general model for clustering binary data. In: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05, 188.
[70] Li Y, Li D, Wang S, Zhai Y. [2014] Incremental entropy-based clustering on categorical data streams with concept drift. Knowledge-Based Systems. 59:33–47.
[71] Li-Na W, Qian L, Yuan Z. [2013] A fuzzy centroids clustering algorithm with between-cluster information for categorical data. Information technology journal. 12:5482-5486.
[72] Mampaey M, Vreeken J. [2013] Summarizing Categorical data by clustering attributes. Data Mining and Knowledge Discovery. 26:130–173.

[73] Mingoti SA, Matos R. [2012] Clustering Algorithms for Categorical Data: A Monte Carlo Study. International Journal of Statistics and Applications. 2:24-32.
[74] Mukhopadhyay A, Maulik U, Bandyopadhyay S. [2009] Multi objective Genetic Algorithm-Based Fuzzy Clustering of Categorical Attributes. IEEE Transactions On Evolutionary Computation. 991-1005.
[75] Ng MK, Wong JC. [2002] Clustering categorical data sets using Tabu search techniques. Pattern Recognition. 35:2783–2790.
[76] Ong K, Li W, Ng W, Lim E. [2004] An Algorithm for Clustering Data streams of categorical attributes. Data warehousing and knowledge discovery Lecture notes in computer science. 3181:209–218.
[77] Park IK, Choi GS. [2015] Rough set approach for clustering categorical data using information-theoretic dependency measure. Information Systems. 48:289–295.
[78] Park SSH, Song JJ, Lee JJH, Lee W, Ree S. [2015] How to measure similarity for multiple categorical data sets. Multimedia Tools and Applications. 74:3489–3505.
[79] Parmar D, Wu T, Blackhurst J. [2007] MMR: An algorithm for clustering categorical data using Rough Set Theory. Data & Knowledge Engineering. 63:879–893.
[80] Qin H, Ma X, Herawan T, Zain JM. [2014] MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. Knowledge-Based Systems. 67:401–411.
[81] Reddy HV, Raju SV, Kumar BS, Jayachandra C. [2014] An Approach for Data Labeling and Concept Drift Detection Based on Entropy Model in Rough Sets for Clustering Categorical Data. Journal of Information & Knowledge Management. 13:1450020.
[82] Rendón E, Abundez I, Arizmendi A, Quiroz EM [2011] Internal versus External cluster validation indexes. International Journal of Computers and Communications 5:27–34
[83] Saha A, Das S. [2015] Categorical fuzzy k-modes clustering with automated feature weight learning. Neurocomputing. 166:422–435.
[84] Saha I, Maulik U. [2014] Incremental learning based multi objective fuzzy clustering for categorical data. Information Sciences. 267:35–57.
[85] Saha I, Sarkar JP, Maulik U. [2015] Ensemble based rough fuzzy clustering for categorical data. Knowledge-Based Systems. 77:114–127.
[86] Seman A, Abu Bakar Z, Mohd. Sapawi A, Othman IR. [2013] A medoids-based method for clustering categorical data. Journal of Artificial Intelligence. 6:257-265.
[87] Seman A, Sapawi AM, Salleh MZ. [2015] Performance Evaluations of κ-Approximate Modal Haplotype Type Algorithms for Clustering Categorical Data. Research Journal of Information Technology. 7:112–120.
[88] Sun Y, Zhu Q, Chen Z. [2002] An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters. 23:875–884.
[89] Szeto LK, Liew AWC, Yan H, Tang S. [2003] Gene expression data clustering and visualization based on a binary hierarchical clustering framework. Journal of Visual Languages & Computing. 14:341–362.
[90] Tamhane AC, Qiu D, Ankenman BE. [2010] A Parametric Mixture Model for Clustering Multivariate Binary Data. Statistical Analysis and Data Mining. 3-19.
[91] Tang Y, Browne RP, McNicholas PD. [2015] Model based clustering of high-dimensional binary data. Computational Statistics & Data Analysis. 87:84–101.
[92] Xiang Z, Islam MZ. [2014] The Performance of Objective Functions for Clustering Categorical Data. Knowledge Management and Acquisition for Smart Systems and Services Lecture notes in computer science. 16-28.
[93] Xiong T, Wang S, Mayers A, Monga E. [2012] DHCC: Divisive hierarchical clustering of categorical data. Data Mining and Knowledge Discovery. 24:103–135.
[94] Xu R, Wunsch II D. [2005] Survey of clustering algorithms. IEEE Transactions on Neural Networks. 16:645–678.
[95] Yang CL, Kuo RJ, Chien CH, Quyen NTP. [2015] Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering. Applied Soft Computing. 30:113–122.
[96] Yang Z, Sun X, Hardin JW. [2012] Confidence intervals for the difference of marginal probabilities in clustered matched-pair binary data. Pharmaceutical Statistics.11:386–393.
[97] Yang Z, Sun X, Hardin JW. [2012] Testing non-inferiority for clustered matched-pair binary data in diagnostic medicine. Computational Statistics & Data Analysis. 56:1301–1320.
[98] Zengyou H, Xiaofei X, Shengchun D. [2002] Squeezer: An Efficient Algorithm for Clustering Categorical Data. J. Comput. Sci. &: Technol. 17:611-624.
[99] Zhang K, Gu X. [2014] An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets. Mathematical Problems in Engineering. 1–8.
[100] Zhang T, Ramakrishnan R, Livny M. [1996] BIRCH: An Efficient Data Clustering Databases Method for Very Large. ACM SIGMOD International Conference on Management of Data. 1:103–11.

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

REVIEW

INTERNET OF THINGS: A REVIEW ON AIDING MACHINERIES, PRACTICES AND SOLICITATIONS

Sameera K, Swarnalatha P*

School of Computer Science and Engineering, VIT University, Vellore, INDIA

ABSTRACT

By accentuating on qualifying mechanics, propriety as well as complications in implementation, this paper provides a summary of the Internet of Things (IoT).The contemporary improvements in various sectors help to emancipate this domain. In order to furnish a contemporary group of applications, the primary assumption would be to integrate exclusive detectors precisely with no personage included. The initial stage of the IoT is the present-day breakaway various mechanizations. The Internet of Things (IoT) is presumed to traverse multiple mechanics in the next few years. This is done by associating substantial devices to one another which helps to validate contemporary implementations. We begin this paper by giving a summary of the Internet of Things. A few practical characteristics of the implementations, proprieties and mechanizations which validate the Internet of Things will be discussed after this summary. A few important issues in this domain from the latest studies are summarized along with an overview on the associated analysis. The correlation between the Internet of Things and other major domains is also portrayed. A collection of scenarios is propounded at the end which can exemplify how various procedures demonstrated in the paper correspond to one another in order to furnish coveted usage of the Internet of Things.

KEY WORDS
Internet of thing;, IoT; security; privacy;

Received: 2 June 2017
Accepted: 18 Aug 2017
Published: 24 Sept 2017

INTRODUCTION

IoT is an area that speaks to the following most energizing mechanical unrest that has been present after the Internet started booming [1-4]. IoT will bring unlimited open doors and effect in each side of our planet. With IoT, we can manufacture shrewd urban communities where parking spots, urban clamor, activity blockage, road lighting, water system, and waste can be observed continuously. We can construct shrewd homes that are sheltered and vitality proficient. Various calamities like tremors, wildfires, etc. can be identified well in advance and contamination of the atmosphere, rivers, lakes and other water bodies can be examined by establishing shrewd frameworks. Production can be transformed by IoT into more biased and shrewder one [5].The sensors empowered by IoT can screen quavers and corporeal circumstances in scaffolds (and additionally structures and recorded landmarks) and give early cautioning that would spare various human lives. IoT will make enormous interruption and advancement in pretty much every industrial section possible. While the IoT offers various energizing possibilities and openings, it stays testing to successfully oversee things to accomplish consistent incorporation of the actual world and the digital one [6-7]. Numerous binding agents of IoT along with network conventions are becoming available. This numeral keeps on expanding. The associating gadgets of IoT which are simple are discouraged even by the plenty of IoT availability conventions along with the binding agents. Same goes with the deciphering of gathered information from these agents and conventions. GSN [8] extends the idea of an implied sensor that is indicated in XML and rendered with a comparing wrapper is given as the primary reflection to creating and interfacing another IoT gadget. Consider the TerraSwarm enterprise [9]. In this, the authors presented a blueprint which was monitor-configured like a fundamental reflection and was actualized using a scripting language. In the Google Fit venture [9], no specific abnormal state reflection is accommodated embodying another gadget sort.

MATERIALS AND METHODS

Understanding the IoT building pieces [10] picks up a superior knowledge into the genuine significance and usefulness of the Internet of Things. [Fig. 1] gives an illustration of various building blocks of IoT. These devices transmit the information in a cycle. First the data is explained clearly including all the constraints. Then each of the devices correlates with each other in order to exchange information. This data will be scrutinized thoroughly and minor changes will be carried out, if necessary. This data is made understandable to the users and an idea is formed eventually. [Fig. 2] illustrates the six fundamental components expected to convey the usefulness of the internet of Things. [Table 1] illustrates the classification of the said components along with their models.

Recognition

IoT needs recognition so as to provide an ID to a resource and couple it with its respective requirement. We can use various techniques for recognition in IoT like uCode [11] and EPC. Besides, tending to the IoT items is basic to separate question identification tag from its location. What's more tending to strategic devices in IoT incorporate internet protocol versions 4 and 6, and internet protocol version 6 over low-power wireless personal area networks [12-13] gives pressure component on to Internet Protocol version 6 rubrics which devise Internet Protocol version 6 tending to proper for remote systems with little capacity. Recognizing article's distinguishing proof and address is basic since ID techniques are not all inclusive extraordinary, so tending to helps to exceptionally distinguish objects. What's more, protests inside the system may utilize open IPs and not private ones. ID techniques are utilized to give an unmistakable character to each protest inside the system.

***Corresponding Author**
Email:
pswarnalatha@vit.ac.in

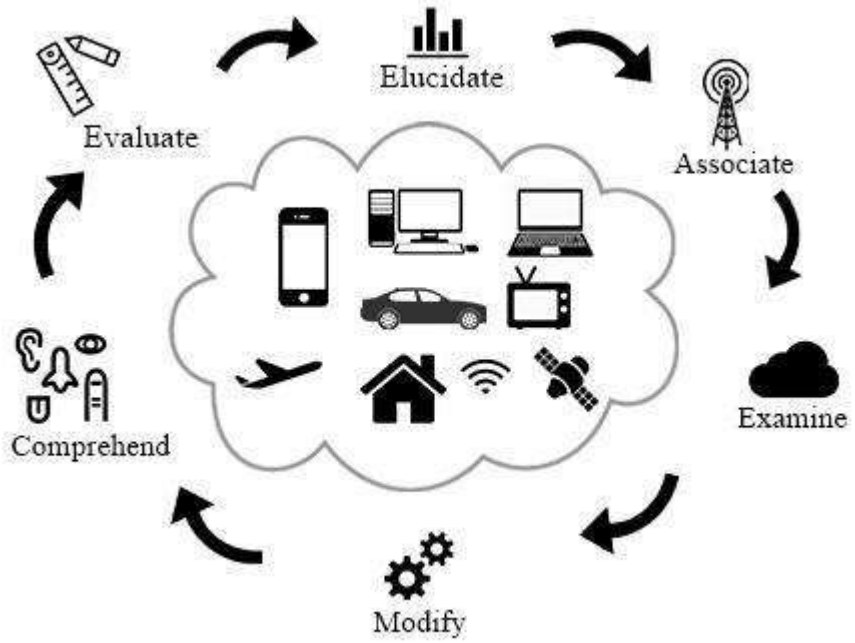


Fig. 1: Some of the devices connected using IoT

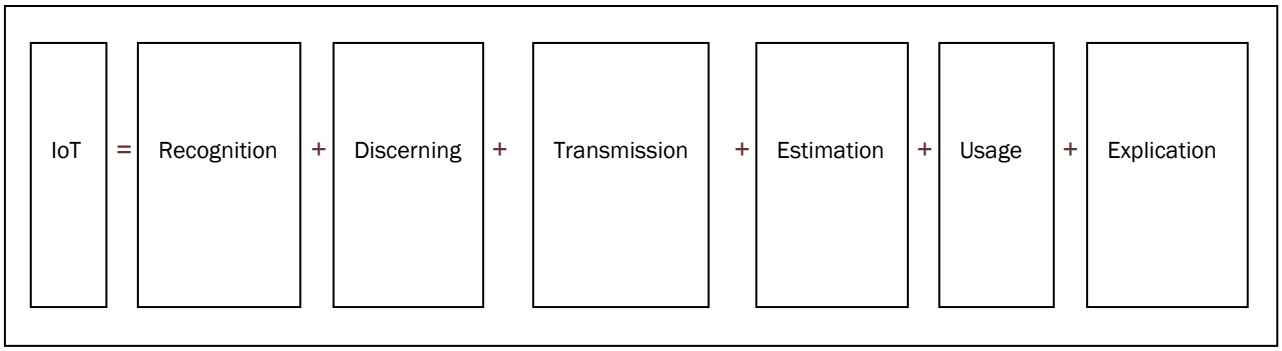


Fig. 2: The IoT essentials

Discerning

Accumulating information from different devices which are in the same system and forwarding it to a storage is called discerning in IoT. Depending on the usage needed, certain undertakings are done by examining the accumulated information. The devices in IoT can be anything like shrewd devices, devices that can be worn and sense fluctuations in human body, actuators, etc. A few firms, for instance, provide various apps which let users to operate various devices using their mobiles [14-16].

To consummate the devices of IoT, reliable serviceability and machines with a sole board are used in general. Detectors and incorporated transmission control protocol along with internet protocol are fixed to these boards. In order to furnish the information needed by the users, these detectors integrate with an intermediate administration egress.

Transmission

In order to provide a certain shrewd usage, the transmission mechanizations in IoT integrate miscellaneous devices with one another. In general, the confluences in IoT must wield by making use of limited capacity when they are present in noisy or lossy transmission connections. Long-Term-Evolution-Advanced, Wireless fidelity, etc. are some of the instances of transmission procedures. The other transmission mechanizations include Radio Frequency Identification, UWB and NFC. The foremost mechanization utilized to accomplish the idea of machine-to-machine was the radio frequency identification. The tag in a radio frequency identification device portrays the identification which presents the originality of a device. A returned pointer is collected by the reader in radio frequency identification

after it sent a query pointer to the tag. In order to recognize a device depending on the return pointers, the database integrates with an operating hub [17]. There are different types of tags in radio frequency identification like dynamic, lethargic, semi dynamic or semi lethargic. Dynamic labels are fueled by battery while latent ones need not bother with battery. Semi-aloof/dynamic labels utilize board control when required.

Table 1: Construction wedges and tools of IoT

IoT Essentials		Models
Recognition	Identifying	EPC, uCode
	Inscribing	Internet Protocol version 4, Internet Protocol version 6
Noticing		Shrewd Devices, Wearable detecting strategies, Implanted instruments, Actuators, RFID label
Transmission		Radio Frequency Identification, Near Field Communication, Ultra-wideband, Wireless Fidelity
Estimation	Hardware	Galileo, Arduino, Raspberry Pi
	Software	Operating system, Cloud
Usage		Originality-associated, Data gathering, Pervasive, Coordinative-concerned
Explication		Web Ontology Language

In case of NFC, a 10cm pertinent scale is needed for transmission among lethargic tags and dynamic readers or two dynamic readers [18]. Wireless fidelity is the transmission mechanization which makes use of wireless signals to transfer information among objects which are less than 100m. away from one another [19]. This mechanization permits brilliant gadgets to impart and trade data without utilizing a switch in some impromptu designs. Depending on Global system for Mobile/Universal Mobile Telecommunications System mechanizations, the LTE was initially designed as a principle wireless transmission of information between mobiles [20]. An enhanced edition of the LTE was released called the LTE-A (LTE Advanced) [21]. It has a built-in add-on for frequency range that maintains up to 100MHz.

Estimation

The heart and soul of the IoT are organizing components and programming implementations. These two provide the data processing capability to the IoT. To execute an IoT implementation, many hardware programs emerged like Raspberry Pi, Arduino, Galileo, etc. Moreover, numerous product stages are used to provide serviceability to IoT. The real-time operating systems are the most crucial here. This is because they operate till the object's stimulation time gets over. The IoT implementations which depend on RTOS can be enhanced using many RTOS. One more crucial element of data processing is cloud storage. It furnishes data storage for the shrewd devices. The users can acquire this information after it is organized in real-time. Many paid and unpaid cloud storage programs and systems are accessible and these can be used by the IoT objects.

Usage

In general, the usage in IoT is classified into four types [22-23]: originality-associated, data-gathering, coordinative-concerned and pervasive. The remaining usage types can make use of originality-associated usage as it is the fundamental type. Data-gathering usage collects and condenses unanalyzed data which requires handling before being outsourced to the IoT implementation. In the sections that ensue, we analyze a few implementations in IoT based on this classification. Reaching the extent of pervasive usage is the eventual objective of every IoT implementation. As there are many complications and issues which need to be tackled, achieving this objective is difficult. Originality-associated and data-gathering usage is furnished by many available implementations. Data-gathering group comprises of shrewd medical management and coordinative-concerned comprises of commercial industrialization and industrial transport systems (ITS).

Individual way of living of users is made simpler and it has become easy to utilize and control gadgets and networks which are connected to one another wirelessly. Building automation systems (BAS) are connected to the web by smart buildings [24]. Detectors and generators like amusement, well-being,

IIOAB JOURNAL
ISSN: 0976-3104

reliability, darkening and brightening, heating, ventilation and air conditioning and so on are used by BAS which enable it to organize and operate various constructing objects. Preservation of constructions and usage of power is also done by BAS.

In order to manage and operate the shipment system, calculation is collaborated with transmission [25-26]. Attaining security, coherence, protection and accessibility of the shipment framework is the key objective of ITS. IoT comprises of four elements: reliability substructure, ITS control unit, location substructure and automobile substructure. These days, integrated automobiles are being termed as crucial because operating them is more secure, amusable as well as systematic. A few examples of this are Audi, Volvo, Google, etc.

People are being less included in building assignments after the industrial automation introduced digital self-operating objects. Four components are to be considered when systems are used to manufacture items easily and precisely: transmission, detection, managing and shipment. In order to operate and manage the performance, usefulness and efficiency of manufacturing systems through the web, commercial industrialization uses IoT.

Detectors and generators are implanted in victims as well as their medications through shrewd medical arrangement. This helps to control, manage and keep track of all the details of the patients' health. Detectors are used in hospitals to gather and examine the patients' data. IoT helps to transfer this examined data to wireless managing units which in turn helps to take appropriate moves. Masimo Radical-7, for instance, controls the victim's position and outlines it to the physicians. Not so long ago, IBM introduced a radio frequency identification mechanization to check if the doctors clean their hands after treating each victim. This technique helps to prevent any unwanted contaminations.

The usage of power in houses and structures is made better by smart grids which use IoT. Energy providers are able to operate the services so that they provide energy equally to the growing number of people through IoT in smart grids. In order to collaborate numerous meters in various constructions to power suppliers, smart grids make use of IoT. The power usage is continuously checked and controlled by these meters. The power suppliers are continuously upgraded by IoT enhancing their resources to satisfy the user demands. Moreover, the rate of incompetency is reduced by smart grids which use IoT. In addition to this, coherence as well as standard of the resources is improved.

One of the examples of pervasive usage is smart city. It makes it simpler for the users to discover the data that attracts them and by this it hopes to enhance the standard of living in cities. Many networks depend on shrewd mechanizations in the smart city scenario which supply the resources needed.

Explication

Considering the current issues and technologies in IoT, explication alludes capacity which concentrates on information shrewdly with various devices that give administrations needed. Learning withdrawal incorporates finding and utilizing assets and demonstrating data. Additionally, it incorporates perceiving and investigating information to settle on feeling of the correct choice to give the correct administration [27]. In this way, explication is portrayed as the heart of the IoT by requesting the correct asset.

EXI is critical with regards to the IoT in light of the fact that it is intended to streamline XML applications for asset obliged conditions. Besides, it lessens data transfer capacity needs without influencing related assets, for example, battery life, code estimate, vitality expended for preparing, and memory measure. EXI changes over XML messages to double to lessen the required transmission capacity and limit the required stockpiling size.

RESULTS

As there are many problems which have to be solved, it had never been simple to understand the innovation of IoT. Confidence, administration, reconcilability, measurability, presentation, adaptability, accessibility are a few among these problems. Implementation coders can apply their resources coherently if these problems are solved. Because of the sensibility of user's confidentiality, all the demands depend on reliability. Moreover, the main task is to analyze the presentation of IoT [28]. The studies portray many of the recognized problems.

For the IoT to be fully developed, many undertakings have to be done even though there have been many studies in this domain. Various analysis programs are being undergone because of the ever-increasing fascination of the administration and business field in this domain. The key problems in this domain are accessibility, framework and presentation. Some examination thinks about have been led in the research centers while others are still in the recreation stage. This is normal since these last difficulties require genuine applications or proving grounds in view of the present advancements; something that has not occurred at an extensive scale yet.

Another beginning IoT inquire about push is to evaluate the system area of brilliant articles to acknowledge new area and setting mindful administrations. The present strategies for area estimation depend on IP.

IIOAB JOURNAL
ISSN: 0976-3104

Named Data Networking (NDN) is among the many prospects which can identify foundation later on the internet [29].

FUTURE DIRECTION

The estimated ascending increase of shrewd objects as well as convergence of affordable framework, association and information establishes the advancement of IoT from a haughty idea to actuality. IoT programs are being extensively utilized due to the escalating emphasis on capacity as well as functioning coherence, extensive association and reduction in object prices [30]. The presumption that IoT creates substantial welfare to enterprises and customers led to this meteoric advancement. This welfare differs with customers, enterprises and administration.

Based on the customers' occupation and location, they can gain individual benefits. Their travelling becomes hassle-free by using associated automobiles which help to steer clear of blocked roads by indicating other ways that they can travel by scrutinizing the traffic updates from other automobiles. In this way, the consumption of power decreases gradually. Wearable objects help to maintain the users' well-being, security and privacy by giving them up-to-date alerts on changes in their bodies or by keeping track of the activities of aged people at home [31].

Surveying the customers' responses helps enterprises to furnish finer usage and manufactured items [32]. The demand for more items and usage can also be discussed. Constructions can be safeguarded by using wireless security, fragile items can be can be made sure to be kept in appropriate prerequisites and movable properties like automobiles and equipment can be made invulnerable with the use of tracing devices and wireless-locking gadgets. Shrewd meters help to reduce scrap in serviceableness and precautionary conservation can be provided by the machine enterprises. Shrewd agriculture helps countrymen to save water by supplying it right when required. The earnings of the enterprises can be improved by enterprise representations depending on the end results but not on the machinery.

IoT can also be of advantage to the administration as well as prominent jurisdictions. Finer wireless equipment can be installed in houses to help senior citizens and to decrease their medical bills. Information can be gathered from numerous automobile users and it can be used to enhance road safety. By turning down the street lights in desolate roads, their productivity can be enhanced.

CONCLUSION

The standard of existence is enhanced by IoT which connects various shrewd objects, mechanizations and implementations. The computerization of everything around us can be made easier by IoT. The latest studies speaking of the various characteristics of IoT, the mechanizations, implementations and conventions that accredit it as well as a survey of the proposition of IoT are set forth in this paper. This paper also brings up certain complications and predicaments which concern the prototype as well as the establishment of IoT applications.

Though there still are many predicaments to be taken care of, the prospects of IoT give the impression of being remarkable. To gain finer incorporation amid the usages of IoT, we put forward the necessity of information-gathering, shrewd unconstrained administration as well as adjustments in convention. Choosing a suitable mechanism is crucial for the favorable outcome of a piece of research, as there are many resemblances among various conventions of IoT. The criteria that have to be fulfilled by IoT mechanization must be expounded meticulously and formulated precisely by the system designers. There are two types of house as well as construction computerization mechanisms: sole-centered and multi-centered. Many features like reliability, refreshment, relaxation, transmission, etc. are collaborated in multi-centered mechanism.

Handling and transfer of data is done by these items. Pliable and compliant mechanisms are integrated to these items. Advanced detectors are maintained by these items and these are used for many implementations which involve reliability, brightening as well as power maintenance and other home automation systems.

CONFLICT OF INTEREST
The authors declare no conflict of interest in relation to the work.

ACKNOWLEDGEMENTS
The majority of this work was done with a lot of help from VIT University, Vellore and the authors are very thankful to the University for contribution.

FINANCIAL DISCLOSURE
None

IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] Raggett D. [2015] The web of things: challenges and oppurtunities. IEEE Ccomputer, 48(5): 26-32.

[2] Want R, Schilit BN, Jenson S. [2015] Enabling the internet of things. IEEE Computer, 48(1): 28-35.

[3] Baresi L, Mottola L, Dustdar S. [2015] Building software for the internet of things. IEEE Internet Computing, 19(2): 6-8.

[4] Atzori L, Iera A, Morabito G. [2010] The internet of things: a survey. Computer Networks, 54(15): 2787-2805.

[5] Ngu HH, Gutierrez M, Metsis V, Nepal S, Sheng Z. [2016] IoT middleware: a survey on issues and enabling technologies. IEEE Internet of Things Journal.

[6] Yao L, Sheng QZ, Dustdar S. [2015] Web-based management of the internet of things. IEEE Internet Computing, 19(4): 60-67.

[7] Qin Y, Sheng QZ, Falkner NJG, Dustdar S, Wang H, Vasilakos AV. [2016] When things matter: a survey on data-centric internet of things. Journal of Network and Computer Applications, 64: 137-153.

[8] Aberer K, Hauswirth M, Salehi A. [2006] A middleware for fast and flexible sensor network deployment. Proceedings of the 32nd International Conference on Very Large Databases, 1199-1202.

[9] Latronico E, Lee E, Lohstroh M, Shaver C, Wasicek A, Weber M. [2015] A vision of swarmlets. IEEE Internet Computing, 19(2): 20-28.

[10] Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M. [2015] Internet of things: a survey on enabling technologies, protocols and applications. IEEE Communications Surveys and Tutorials

[11] [2015] Google Fit. https://developers.google.com/fit/.

[12] Koshizuka N, Sakamura K. [2010] Ubiquitous ID: standard for ubiquitous computing and the internet of things. IEEE Pervasive Computing, 9: 98-101.

[13] Kushalnagar N, Montenegro G, Schumacher C. [2007] IPv6 over low-power wireless personal area networks (6LoWPANs): overview, assumptions, problem statement and goals. RFC4919, 10.

[14] Montenegro G, Kushalnagar N, Hui J, Culler D. [2007] Transmission of IPv6 packets over IEEE 802.15.4 networks. Internet Proposed standard RFC 4944.

[15] Pilkington K. [2014] Revolv teams up with home depot to keep your house connected. CNET-News.

[16] [2014] SmartThings. Home Automation, Home Security and Peace of Mind.

[17] Rushden U. [2012] Belkin brings your home to your fingertips with WeMo home automation system. Press room Belkin.

[18] Want R. [2006] An introduction to RFID technology. IEEE Pervasive Computing, 5: 25-33.

[19] Want R. [2011] Near field communication. IEEE Pervasive Computing, 10: 4-7.

[20] Ferro E, Potorti F. [2005] Bluetooth and Wi-Fi wireless protocols: a survey and a comparison. IEEE Wireless Communications, 12: 12-26.

[21] Crosby GV, Vafa F. [2013] Wireless sensor networks and LTE - a network convergence. IEEE 38th Conference on Local Computer Networks (LSN), 731-734.

[22] Ghosh A, Ratasuk R, Mondal B, Mangalvedhe N, Thomas T. [2010] LTE-advanced: next generation wireless broadband technology. IEEE Wireless communications, 17: 10-22.

[23] Xiaojiang X, Jianli W, Mingdong L. [2010] Services and key technologies of the internet of things. ZTE Communications, 2: 011.

[24] Gigli M, Koo S. [2011] Internet of things: services and applications categorization. Advances in Internet of Things, 1: 27-31.

[25] Finch E. [2001] Is IP everywhere the way ahead for building automation? Facilities 19: 396-403.

[26] Talcott C. [2008] Cyber-physical systems and events. Software Intensive Systems and New Computing Paradigms. Wirsing M, Banatre J, Holzl M, Rauschmayer A. Springer Science and Business Media 101-115.

[27] Yongfu L, Dihua S, Weining L, Xuebo Z. [2012] A service-oriented architecture for the transportation cyber-physical systems. 31st Chinese Control Conference, 7674-7678.

[28] Barnaghi P, Wang W, Henson C, Taylor K. [2012] Semantics for the internet of the things: early progress and back to the future. International Journal on Semantic Web and Information Systems, 8: 1-21.

[29] Uckelmann D. [2012] Performance measurement and cost benefit analysis for RFID and internet of things implementations in logistics. Qualifying the Value of RFID and the EPCglobal Architecture Framework in Logistics, Springer, 71-100.

[30] Zhang L, Afanasyev A, Burke J, Jacobson V, Crowley P, Papodopoulos.

[31] Cognizant Report Reaping the Benefits of the Internet of Things. Retrieved from http://www.cognizant.com/InsightWhitePapers/Reaping-the-Benefits-of-the-Internet-of-Things.pdf.

[32] Pundir Y, Sharma N, Singh Y. [2016] Internet of things (IoT): challenges and future directions. International Journal of Advanced Research in Computer and Communication Engineering, 5(3): 960-964.

[33] Davies R. [2015] The internet of things oppurtunities and challenges. European Parliamentary Research Service. Retrieved from www.europarl.europa.eu/RegData/.../EPRS_BRI(2015)557012_EN.pdf.

ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

ENHANCED HADOOP PERFORMANCE ANALYSIS USING HADOOP ECO-SYSTEM

Shivam Suryawanshi[1], Jabanjalin Hilda[2*]

[1]*School of Computer Science and Engineering, Vellore Institute of Technology - Vellore, T.N, INDIA*

[2]*Faculty of School of Computer Science and Engineering, Vellore Institute of Technology - Vellore, T.N, INDIA*

ABSTRACT

All over the World, the idea of Big Data analytics is constantly growing in the software domain. Hadoop as an open source has become a popular Java framework for large scale data processing in recent years. Day by day, the rate of data is also increasing through different sources, and to analyze such huge data it is required to be processed on Hadoop Distributed File System (HDFS) rather than processing using traditional methods. The Apache Hadoop eco-system supports many open source tools for analyzing substantial datasets. Three well-known tools for data analyzing in the HDFS are Hive, Pig, and MapReduce. In this project, we are analyzing crime dataset which is imported on HDFS using Sqoop tool and further, it is analyzed using Hive, Pig, and MapReduce program. The result shows that it is an enhanced Hadoop performance analysis though Hadoop eco-system components where Hive works more efficiently than Pig and MapReduce Program. Later, results are compared with MapReduce program. In this crime data analysis, Hive outperforms 2.23 times than MapReduce and 1.90 times than Pig.

KEY WORDS
HDFS, Hadoop, MapReduce, Hive, Pig, Sqoop

INTRODUCTION

These days a lot of information are coming from various sources like Social network profiles, advanced media - audio, photos, and web sources, so on. The successful storage, querying and analyzing of these information has turned into a challenging test to the business. When it comes to crime, size of crime data growing more day by day and storing, analyzing, processing such large data has dependably been a big concern in the field of database management domain. Nowadays big data has become a part of handling such issues related to large information. There are a few inquiries with regards to crime. Questions like - Is crime a serious issue where you live? What sorts of crimes happen frequently? Is the crime rate increasing or decreasing where you live? Do you think the world will be secure or riskier later on? [1] And many more. To solve this problem, it is required to analyze the crime related data firstly using big data technology. It is also important to characterize certain definitions that are identified with Big Data and Hadoop.

Big Data

Received: 4 June 2017
Accepted: 30 Sept 2017
Published: 12 Oct 2017

Big Data are huge-volume, high-speed, as well as huge-variety data resources that require new types of handling to ensure upgraded process enhancement [2]. Increased processing speed, storage limit, and systems administration have made information to develop in every one of the 4 measurements. The four characterizing qualities of Big Data- volume, variety, value and velocity - are incorporated for the better performance of data handling. There are distinctive methods for characterizing and comparing Big Data with the customary information, for example, information size, content, collection and processing. Huge information has been characterized as expansive data sets that can't be prepared using conventional handling strategies, for example, Relational Database Management Systems, in an average preparing time. So for that new technology comes under big data is nothing but Hadoop [3], [4].

Hadoop

Hadoop is the system which allows to store and process a huge amount of data across multiple computers connected in distributed environment. By using the Hadoop technology, we can scale up from single-node cluster to multi-node cluster of machines, each offering different storage capacity and computation. Hadoop can be used in a different application in order to process the data as the data is generating more and more information on a daily basis, and it is becoming very difficult to handle the data. The importance of big data technologies is providing more accurate analysis, which can be used for decision-making in any business process. There are different big data technologies such as operational Big Data which includes a system like MongoDB where data is primarily stored [5], [20]. These systems are supposed to take an advantage of new distributed computing systems that have created over the earlier decade which will run productively. This tends to information workloads significantly less demanding to manage, simple, quicker, and less costly to execute. Another Big data technology is, which uses the parallel environment such as MapReduce programming that provides analytical capabilities to review and complex examination analysis all of the type of the data. We can use MapReduce to scale up the single machine to multiple machines. It provides different method for mapping the information which is integral to the capacities gave by SQL.
Using mapping function, it takes the data from the huge dataset and distributes to multiple machines. [Fig. 1] Shows the Hadoop Framework which incorporates MapReduce layer and HDFS layer. A little Hadoop group incorporates a master part and different slave part. The framework consists of a DataNode, NameNode, Job Tracker and Task Tracker [6].

***Corresponding Author**
Email: jabanjalin.hilda@vit.ac.in
Tel.: +91-7598193077

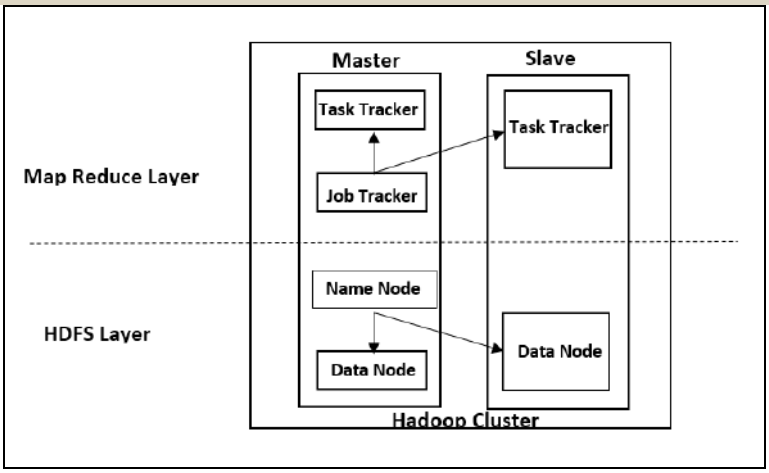


Fig. 1: Hadoop Framework [5]

- ***Data Node:*** Constantly inquire as to whether there is something for them to do, mainly used to track which data nodes are up or and which data nodes are down.
- ***Name node:*** Manages the record framework name space, it monitors where each block is present.
- ***Job tracker:*** Assigns the mapper job to undertaking tracker nodes that have the information or are near the information (same block)
- ***Task tracker:*** Keep the work as near the information as could be expected under the circumstances.

NameNode stores MetaData about the information being put away in DataNodes though the DataNode stores the real Data. JobTracker is an expert which makes and runs the jobs. JobTracker which can keep running on the NameNode distributes the job to TaskTrackers which keeps running on DataNodes; TaskTrackers run the assignments and report the status of the job to JobTracker. A slave part is responsible tracking job with the help of DataNode and TaskTracker. In a single-node cluster, both the NameNode and DataNode use the same machine for processing the data. In a multimode cluster, NameNode and DataNodes are ordinarily on various machines. There is one and only NameNode in a bunch and numerous DataNodes [6].

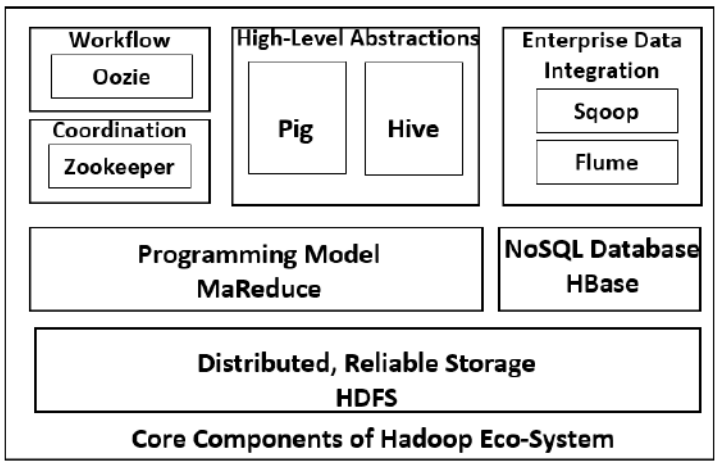


Fig. 2: Hadoop Eco-system components [8]

In addition, by enhancing the Hadoop we can reduce processing time, data size to read and other parameters in Hadoop MapReduce environment [7], [8]. This paper focuses on conceptual technologies, tools about huge information analysis and results shown through a couple of situations how it's useful for associations inside different segments if examination are conducted effectively. There are core components of the Hadoop eco-system, they are Sqoop, Pig, Hive, and Map Reduce. [Fig. 2] shows the architecture of Hadoop Eco-system components [9].

LITERATURE SURVEY

At current situation, each organization is confronting common difficulties which should be adapted up rapidly and effectively. Hadoop was incorporated with various segments that can be utilized for placing, executing and analyzing information significantly as well as proficiently. The decision of a specific utility relies on upon the necessities of the data analysis, the user technical knowledge, and the tradeoff between development time and execution time. Big data environment exhibits extraordinary turn for

organizations inside different parts to compete an upper hand. There are irregularities and difficulties inside Big Data analysis: adequate calculations to cover raw information or investigation, dependability and integrities of Big Data, information storage issues and MapReduce paradigm [5].

Fuad et al. [9], presented a conference paper which introduces the execution time of Hive, Pig, and MySQL Cluster on a basic information system with basic queries while the information is developing. In this, they have framed MySql database issue related to storing and processing information. The issue with MySQL Cluster is that as the information becomes bigger, amount of time to prepare the information increments and for that extra sources might be required. With Hadoop - Hive and Pig, handling time can be speedier than MySQL Cluster. Hence, three data analyzers with similar information model will run basic queries and to discover at what number of columns Hive or Pig is speedier than MySQL Cluster. They have worked on Group-Lens data set where an outcome shows that Hive is the most proper for this information model in a minimal effort hardware condition.

Prabhu et al. [10], presented a paper and they have taken web log information for the experiment and probed on Native Hadoop attributes that is a benchmark framework where they examined from the outcome i.e. when they streamline Hadoop framework attributes then they can enhance the framework execution. In this way they worked on enhancing the parameter in view of the framework assets and application and they also talked about why Hadoop setup must be transformed from its default to particular framework. After executing the experiment, they noticed that native execution has enhanced by 32.97%. For that they have considered couple of parameters.

Sathyadevan et al. [11], presented a paper where they explained about, crime data analysis and its prevention methods which can be a precise way of recognizing the common crime patterns. In their approach, they are predicting the areas where crimes are happening for more number of time and they can imagine crime inclined zones. In this paper they have used the idea of data mining where they are extracting the unknown existing features, valuable data from an unstructured node. Here they gone through an approach between software engineering and criminal equity to build up an information mining methodology that can help tackle crimes speedier. Rather than concentrating on reasons for crime event like criminal foundation of wrongdoer, political hatred etc., they are concentrating chiefly on crime variables of every day.

Alshammari et al. [12], presented a paper where they displayed Enhanced Hadoop (H2Hadoop) that permits a Name-Node to distinguish the blocks inside the cluster where particular data is present. In H2hadoop, input data is less, also input functions are lessened by the quantity of Data-Nodes conveying the main data-blocks that are again distinguished by sending a job to Task-Tracker. They have added some control feature on Name-Node whose purpose is to assign a task of particular data to a Data-Node without sending it to a whole cluster. They talked about the proposed work of H2Hadoop and showed the execution time of H2Hadoop which is efficient with respect to native Hadoop.

HADOOP ECOSYSTEM AND ITS METHODOLOGY

Problem Evaluation

With persistently expanding population, crimes and its rate dissecting related information is a tremendous issue for governments to settle on vital choices to keep up the peace. This is truly important to guard the residents of the nation from violations. The best place to admire opportunity to get better is the voluminous raw information that is created all the time from different sources by applying Big Data Analytics which breaks down to specific patterns that must be found, so that law can be kept up legitimately and there is a faith of security and prosperity among the people of the nation. In this paper, the three methodologies are differentiated with the help of a use case for Hadoop: Crime data investigation [13], [14].

The dataset examined in these tests were produced by a MapReduce program, using these crime data set as info, it is possible to check the output of the executed programs for precision. The issue is characterized further in the following segment, trailed by areas on the Hive, Pig and MapReduce solutions, and then the outcomes. Three well known tools for data analyzing occupant in the HDFS are Hive, Pig, and MapReduce. Hive gives a SQL like front end with a database foundation. Pig gives high level programming language to perform information processing that additionally empowers the users to misuse the parallelism innate in a Hadoop Cluster. MapReduce needs a PC program (frequently Java Programming) for inserting, handling and showing the output information. Hive and Pig produce MapReduce code to do the genuine performance analysis [15].

Sqoop

Sqoop is a command line tool used to transfer data from RDBMS to HDFS and vice versa [16]. Firstly, user can import data via sqoop from RDBMS (either MySql, SQL Server, PostgreSQL, etc.). After importing data

from RDBMS it will sink to HDFS using Hadoop MapReduce functionality. Following workflow shows Sqoop Architecture, [Fig. 3].

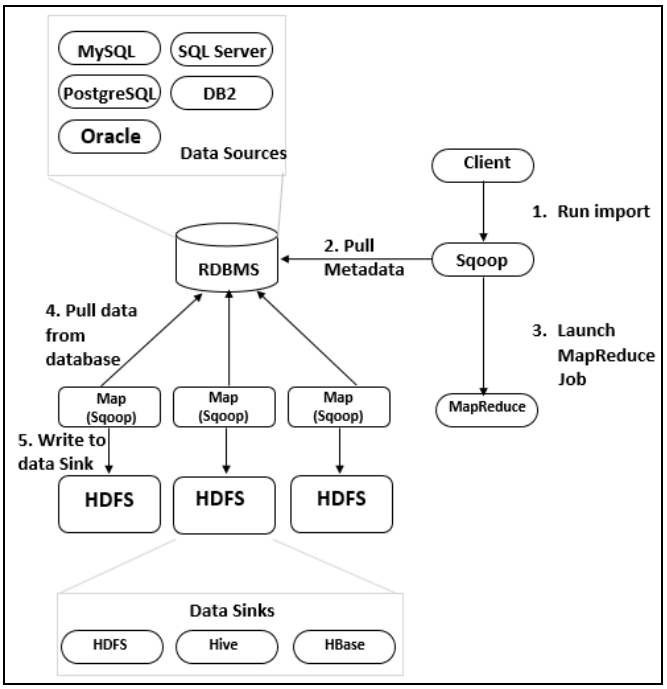


Fig. 3: Sqoop workflow architecture. [16]

HIVE

The Apache Hive device is not a RDBMS tool, it is a part of the Hadoop eco-system, which works on the data which is stored in HDFS using HiveQL (HQL), is a Structured Query Language (SQL) interface, to solve or execute the query based on the available data [17]. This SQL based language in Hadoop domain gives good platform to view the information present in tables. Hive makes a query plan that implements the HQL in a progression of MapReduce projects, produces the code for these projects, after that executes the code, gives appropriate results. Following structure [Fig. 4] is the HIVE Workflow architecture. It shows that, how hive and Hadoop works together when it comes to MapReduce task.

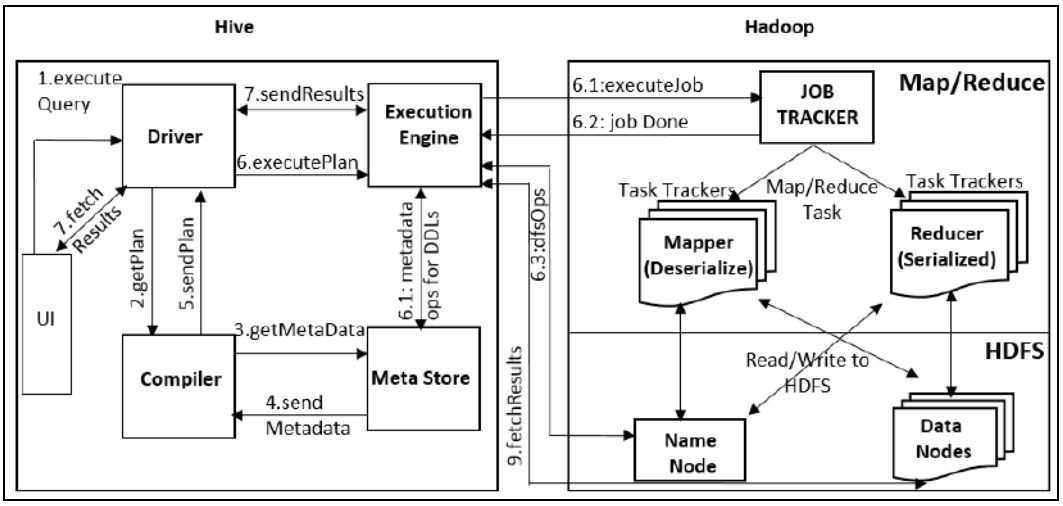


Fig. 4: Hive workflow diagram. [17]

PIG

Apache Pig tool is another information analysis device in the Hadoop Eco-system [18]. Pig is a data flow language, Pig Latin, which allows the client to determine joins, and different calculations without the need to compose an entire MapReduce program. Like Hive, Pig creates a flow of MapReduce projects to solve the data analysis steps. Following flow chart [Fig. 5] will describe PIG architecture.

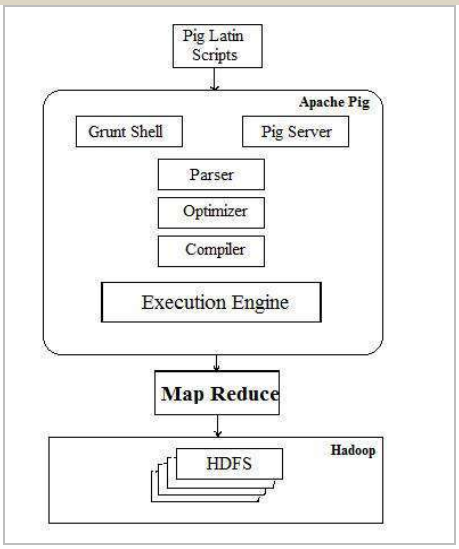


Fig. 5: Pig architecture. [18]

Map reduce

MapReduce is a system utilized by Google for handling large measures of information in a distributed domain and Hadoop is Apache's open source execution of the MapReduce structure. Hadoop is helpful for putting vast volume of information into Hadoop Distributed File System and that information get prepared by MapReduce paradigm in parallel. MapReduce is a versatile and effective programming model to perform substantial scale information. At the point when handling this monstrous information asset has been constrained to single PCs, computational force and capacity rapidly get to be bottlenecks. This massive amount of information can be handled in distributed environment by processing each task one by one. The Hadoop MapReduce structure gives a stage to such parallelization of tasks [19].

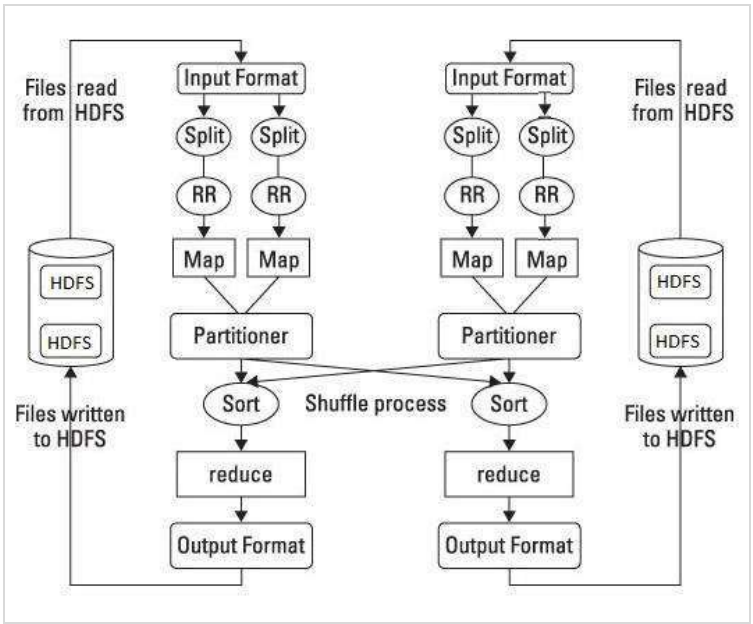


Fig. 6: Mapreduce paradigm. [19]

MapReduce Code Snippet

```
map(LongWritable k,Text v,Context c)
String s = v.toString();
        String s1[]=s.split(",");
        String type_crime=s1[6];
        c.write(new Text(type_crime), new IntWritable(1));
String report=s1[2];
        c.write(new Text(report), new IntWritable(1));
```

```
reduce(Text k,Iterable<IntWritable> v,Context c)
int count=0;
        while(v.iterator().hasNext())
                IntWritable i=v.iterator().next();
                count+=i.get();
        c.write(k,new IntWritable(count));
```

SIMULATION AND EXPERIMENTAL RESULTS

We have setup an experiment on Hadoop Cluster in Linux based Cent OS with single node running and configuration as: Minimum 2 GB RAM and Minimum 100GB Hard Disk Space. In this Crime Data Analysis, we are focusing on following three problem statements.

Table 1: Problem Definition

Problem #	Problem Statement	Result
Problem 1	Finding total number of crimes reported by	Fig. 7.1
Problem 2	Finding total number of Each Crime in last 6 Years (2011-2016)	Fig. 7.2
Problem 3	0.29 sec	Fig. 7.3

Using Sqoop Tool we are storing MySql data to HDFS. It can be done via following command. We can also revert the process i.e. HDFS to MySql [16].
Sqoop *->sqoop import --connect jdbc:mysql://localhost/training --username training --password training --table sr --m 1 --target-dir Crime_1;*
By using above statement, we can fetch the data present on the HDFS, Stored data further can be analyzed by hive and pig, results of each problem is summarized in the following section.

Hive analysis

Hive information is placed in HDFS, which is additionally sent to different nodes. This information is placed in a plain document with CSV, as provided by Sqoop. Hive will read entire file using indexing which results to faster query output. Hive won't execute a MapReduce Task if it does not include either of join, group by, order by aggregate operation. By querying this operation, hive can promptly begin the MapReduce task, which may requires 5-10 seconds to begin the MapReduce. [Table 2] shows Hive analysis. Solution using Hive for -

- ***Problem 1:*** *hive> select report_by,count(*)as tot from cdata group by report_by order by tot desc;*
- ***Problem 2:*** *hive> select" TOTAL CRIME FOR LAST 8 YEARS ", crime_type,count(*)as tot from cdata group by crime_type order by tot desc ;*
- ***Problem 3:*** *hive> hive> select year,count(*)as tot from cdata group by year order by tot desc;*

Table 2: Hive result analysis

Tool Hive:	Starting Time	Finishing TIme	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14.31:02	14.31.59	28.74	2	1	Successful
Problem 2	15:02:21	15:02:34	13.01	2	1	Successful
Problem 3	15.43:26	15:43:47	21.30	2	1	Successful

Pig analysis

Pig performs well with huge size of data. Pig executes a well ordered approach as characterized by the developer. If the query given by the developer is not a complex one (query which is included with joins and sorts) then Pig will not work properly. Pig solves every step one by one, which can expend more time in this case. Whenever information need composite job and more joining operation then Pig can deal with it productively by processing every level and persistently processing the next levels. Pig uses Grunt Shell to execute its task. [Table 3] shows pig analysis which is done after executing following script. Solution using Pig, for –

```
Problem 1: grunt> cri1 = LOAD '/user/training/cri ' using PigStorage(',') AS
(mon:int,year:int,report_by:chararray,loc:chararray,losa_code:chararray,losa_name:chararray,type_crime:
chararray);
grunt> cri2 = FOREACH cri1 GENERATE report_by;
grunt> by_loc = group cri2 by report_by;
grunt> count_crimes = foreach by_loc generate group as loc,COUNT(cri2) as TOT;
grant> cri6 = order count_crimes by TOT desc;
grunt> dump cri6;
```

Problem 2: *grunt> cri1 = LOAD ' /user/training/cri ' using PigStorage(',') AS (mon:int,year:int,report_by:chararray,loc:chararray,losa_code:chararray,losa_name:chararray,type_crime: chararray);*
grunt> cri2 = FOREACH cri1 GENERATE type_crime;
grunt> by_type_crime = group cri2 by type_crime;
grunt> describe by_type_crime ;
grunt> count_crimes = foreach by_type_crime generate group as type_crime,COUNT(cri2) as TOT;
grant> cri6 = order count_crimes by TOT desc;
grunt> dump cri6;

Problem 3: *grunt> cri2 = FOREACH cri1 GENERATE year;*
grunt> cri3 = group cri2 BY year ;
grunt> cri5 = foreach cri3 generate group as year,COUNT(cri2.year) as TOT;
grunt> cri6 = order cri5 by TOT desc;
grant>dump cri6;

Table 3: Pig Result Analysis

Tool Pig:	Starting Time	Finishing TIme	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14.47:19	14.47.59	33	2	1	Successful
Problem 2	15:10:51	15:11:34	15.56	2	1	Successful
Problem 3	15.52:06	15:52:47	24	2	1	Successful

Map reduce analysis

Here CrimeDriver class contain Mapper and Reducer methods. After performing a following command, it will start execution where mapping, shuffling and reduce process will happen. [Table 4] shows result for three given problem statements using MapReduce Program. And [Fig. 6] gives MapReduce paradigm. In MapReduce scenario, when the database is compiled for the given problem statements 1,2,3 it requires 36,19 and 30 seconds respectively where number of mappers are 2 and reducer is 1.
[training@localhost workspace]$ hadoop jar crime.jar CrimeDriver cri MROUT1]

Table 4: MapReduce Result Analysis

Tool MapReduce:	Starting Time	Finishing Time	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14:22:14	14.22.50	36	2	1	Successful
Problem 2	15.20.24	15.20.43	19	2	1	Successful
Problem 3	15:30:42	15:31:12	30	2	1	Successful

The given problem statements are used to analyze the Crime Data and results of those are shown on above charts and graph using R tool. In first problem statement, we are finding total number of crimes reported by, result is shown in figure [Fig.7(1)]. Cambridge shire Constabulary 454435 (51.3%), City of London Police 41745 (4.3%), Durham Constabulary 390220(44%). In second problem statement, we are finding total number of Each Crime in last 6 Years (2011-2016), result is shown in figure [Fig. 7(2)], where Anti-social behavior 357187 counts more number of crimes from year 2001 to 2016. In third problem statement, we are finding total number of crimes per Year. In this case, total number of crimes are counted for year 2011 to 2016 in which crimes happened in year 2015 is close to 20000 which is again more. Result is shown in figure [7(3)]. Following part will show [Fig. 7(1)], [Fig. 7(2)], and [Fig. 7(3)] and also gives us an analysis using each of the Hadoop eco-system.

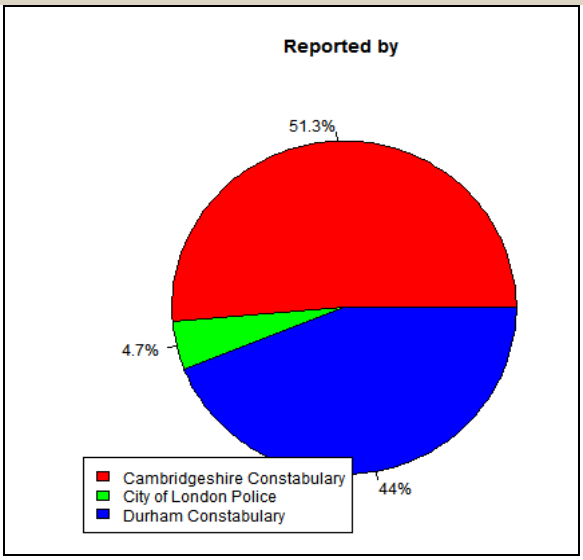


Fig. 7(1): Result for Problem #1

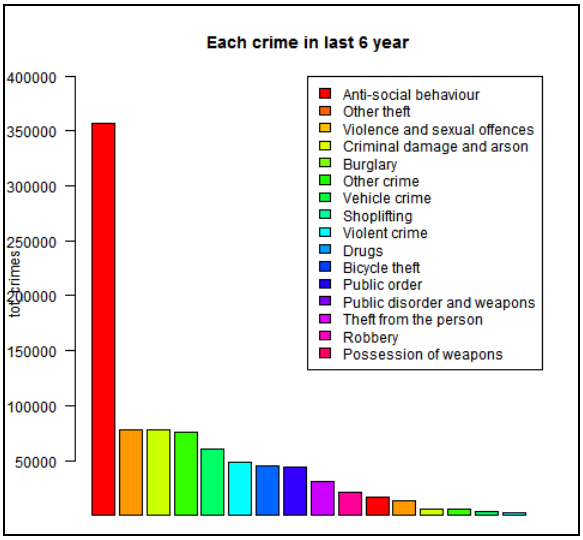


Fig. 7(2): Result for Problem #2

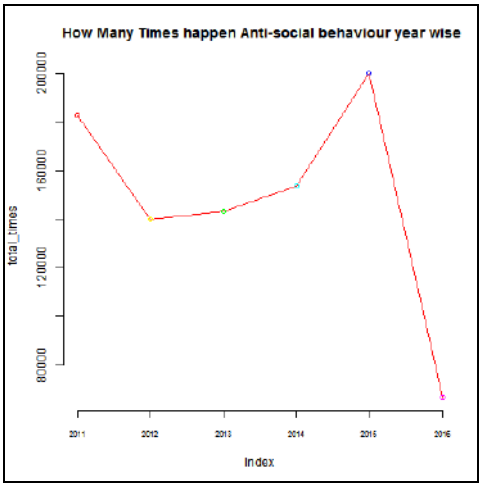


Fig. 7(3): Result for problem #3.

[Fig. 8] shows a graph where, we have analyzed the performance for each of the Hadoop Eco-system tool – Hive, Pig, MapReduce. Results of the graph shows that Hive works efficiently as compared to Pig and

MapReduce for the given use case scenario: Crime Data Analysis. Similarly, we can analyze different problems related to crime data and results of this analysis which can be helpful for restricting crimes in future. If we consider average execution time for Hive, Pig and MapReduce, we can conclude that Hive is 2.23 times faster than MapReduce and 1.90 times than Pig for the given scenario. Also, average line of code used by Hive and Pig are very lesser than MapReduce Program. [Table 5] describes the comparison of each tool with respect to average time complexity and number of code lines.

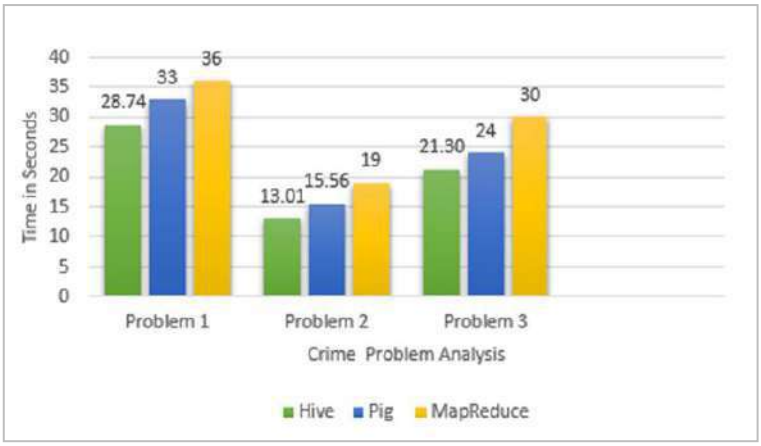


Fig. 7(4): Enhanced hadoop performance analysis.

Table 5: Hive, Pig, MapReduce Performance Comparison

Tool	Average Total Time Taken (seconds)	Time Relative to MapReduce	Line of Code Usage
Hive	38.05	2.23 times	2-4
Pig	72.56	1.17 times	10-12
MapReduce	85	1 .0 times	70-100

CONCLUSION

This paper presents data analysis of huge dataset utilizing three unique things that are a piece of the Hadoop environment – Hive, Pig and MapReduce. The application presented here is a crime Data investigation. The issue is clarified top to bottom and after the simulation, results are shown for the three tools. Complete dataset is accessible from https://data.police.uk/data/ and after successful operations, additionally the R techniques used to display the analysis and plot the outcomes. Results are appeared for each of the three tools with 8lacs set of records. Results show that Hive is more efficient when compared to Pig and MapReduce which shows that it is enhancing Hadoop Performance for huge data set. Hive is 2.23 times faster than MapReduce and 1.90 times than Pig. Also, line of code used by Hive and Pig are very lesser than MapReduce Program. In the future part, focus should be on finding the different parameters which can minimize the Hadoop performance for large size of data using different strategies.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
I express my sincere gratitude to Prof. Jabanjalin Hilda, Senior Assistant Professor in VIT University - Vellore, Department of Software Systems (SCOPE), for her expert guidance and invaluable support in Big Data – Hadoop Technology. I am extremely thankful to VIT University – Vellore for providing the infrastructure facilities for this project and also thankful to my Departmental Dean, HOD, all the Staff Members and my batch mates for their encouragement throughout the course of present work. Finally, I would like to acknowledge with gratitude, the support, and love of my family – my parents, friends without them it would not have been possible.

FINANCIAL DISCLOSURE
None

REFERENCES

[1] "Introduction to Crime Data Analysis", Developed by Garner Clancey.

[2] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey", Journal of Big Data 2015.

[3] Pual C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, "Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill Osborne Media ©2011 book.
[4] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, et al., "Big data and its technical challenges," Communications of the ACM, vol. 57, pp. 86-94, 2014.
[5] S Vemula. "Hadoop Image Processing Framework."
[6] http://hadoop.apache.org
[7] T. White, Hadoop: The definitive guide: "O'Reilly Media, Inc.", 2012.
[8] Herodotou H.[2011] Hadoop performance models". arXiv preprint arXiv:1106.0940
[9] Ammar Fuad, Alva Erwin, Heru Purnomo Ipung, [2014]Processing Performance on Apache Pig, Apache Hive and MySQL Cluster", International Conference on Information, Communication Technology and System,
[10] Swathi Prabhu, Anisha P Rodrigues, Guru Prasad MS & Nagesh HR.[2015] Performance Enhancement of Hadoop MapReduce Framework for Analyzing BigData, Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on March 2015.
[11] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S, "Crime Analysis and Prediction Using Data Mining", Networks & Soft Computing (ICNSC), 2014 First International Conference on August 2014.
[12] Hamoud Alshammari, Jeongkyu Lee and Hassan Bajwa, "H2Hadoop: Improving Hadoop Performance using the Metadata of Related Jobs", IEEE TRANSACTIONS ON Cloud Computing 2015.
[13] Rachel Boba "Introductory Guide to Crime Analysis and Mapping", November 2001 Report to the Office of Community Oriented Policing Services.
[14] Arushi Jaina, Vishal Bhatnagara, [2015] Crime Data Analysis Using Pig with Hadoop", International Conference on Information Security & Privacy (ICISP2015), 11-12, Nagpur, INDIA
[15] Prachi Pandey, Sanjay Silakari, Uday Chourasia,[2016]A Comparative Study of Hadoop Family Tools", International Journal of Computer Science and Information Technologies, 7(3): 1620-1623
[16] http://sqoop.apache.org/docs/1.4.0-incubating/SqoopUserGuide.html
[17] http://hive.apache.org
[18] http://pig.apache.org
[19] Lu Jiamin, Feng Jun.[2015] A Survey of MapReduce based Parallel Processing Technologies", Big Data, Cloud & Mobile Computing,China Communication, Vol 11
[20] Yunquan Zhang, Ting Cao, Shigang Li, Xinhui Tian, Liang Yuan, Haipeng Jia, and Athanasios V. Vasilakos,[2016] Parallel Processing Systems for Big Data: A Survey, Proceedings of the IEEE 104(11).