IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 6 Issue 6

June 2015

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 6 Issue 6 June 2015
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Intrusion Detection and Countermeasure of Virtual Cloud Systems - State of the Art and Current Challenges

Andrew Carlin, Mohammad Hammoudeh
School of Computing, Mathematics & Digital Technology
Manchester Metropolitan University
Manchester, UK

Omar Aldabbas
Al-Balqa Applied University
Faculty of Engineering
Salt, Jordan

***Abstract*—Clouds are distributed Internet-based platforms that provide highly resilient and scalable environments to be used by enterprises in a multitude of ways. Cloud computing offers enterprises technology innovation that business leaders and IT infrastructure managers can choose to apply based on how and to what extent it helps them fulfil their business requirements. It is crucial that all technical consultants have a rigorous understanding of the ramifications of cloud computing as its influence is likely to spread the complete IT landscape. Security is one of the major concerns that is of practical interest to decision makers when they are making critical strategic operational decisions. Distributed Denial of Service (DDoS) attacks are becoming more frequent and effective over the past few years, since the widely publicised DDoS attacks on the financial services industry that came to light in September and October 2012 and resurfaced in the past two years. In this paper, we introduce advanced cloud security technologies and practices as a series of concepts and technology architectures, from an industry-centric point of view. This is followed by classification of intrusion detection and prevention mechanisms that can be part of an overall strategy to help understand identify and mitigate potential DDoS attacks on business networks. The paper establishes solid coverage of security issues related to DDoS and virtualisation with a focus on structure, clarity, and well-defined blocks for mainstream cloud computing security solutions and platforms. In doing so, we aim to provide industry technologists, who may not be necessarily cloud or security experts, with an effective tool to help them understand the security implications associated with cloud adoption in their transition towards more knowledge-based systems.**

Keywords—Cloud Computing Security; Distributed Denial of Service; Intrusion Detection; Intrusion Prevention; Virtualisation

I. Introduction

Cloud computing is a growing facet of the technical infrastructure of modern information systems. It provides a way to deliver the demand of users for near consistent access to their data and software resources regardless of their physical position [1]. Many industries have already adopted cloud computing given the benefits, such as the elasticity, agility, adaptability and availability, that it brings to the Information Technology (IT) infrastructure [2]. The cloud model reduces industrial costs by simplifying the process of installing hardware and software updates and ensuring availability and adaptability of computing resources as required. These properties allow resources to be deployed as necessary to manage peak capacity or to support prolonged industrial growth. Cloud computing allows a rapid response to these demands when compared to traditional IT models, where resources has to be managed and installed on-premise causing both high start-up and maintenance costs. In the cloud model, resources are rented, often autonomously, as required from the Cloud Service Provider (CSP). Equally, resources can be returned to the ‘pool’ when not being used leading to greater resource utilisation efficiency, lower cost and ‘greener’ computing [3]. This flexible infrastructure allows industries to focus on their own business processes, while the computing elements are managed by a CSP rather than by the company’s own IT departments. Some companies choose to make use of internal private clouds allowing them to dynamically deploy their own computing resources as appropriate.

The cloud computing model relies on maintaining a certain level of trust between clients and providers to ensure that client’s data is secure and that an agreed Quality of Service (QoS) is provided at all times. There is also trust from the providers to the clients to avoid malicious activity against the provider either through direct, e.g., insider attacks, or indirect means, e.g., security flaws in applications that are deployed on the cloud. Changes in the user’s use of resources is monitored by the CSP. Unusual discrepancies in this use can affect trust in the user and therefore affect the services they receive [4].

One issue that has hampered the uptake of cloud computing by businesses is the issue of security. This covers the security concerns of all stakeholders from end users, to clients and to the CSPs themselves. The relationships between clients and CSPs are underlined by service agreements, which define the service that clients should receive. These agreements define the responsibilities of all parties with regards to accessibility, data integrity, confidentiality and security.

The power of the cloud is a tempting target for exploitation from attackers aiming to launch further attacks. In 2011, a hacker used Amazon’s Elastic Computing Cloud Service (EC2) to attack Sony’s online entertainment systems, compromising more than 100 million customer accounts. This was the largest data breach in U.S. history [5]. A similar attack was later used to prevent users from logging on to Sony services [6]. Another example of the power of the cloud being utilised for malicious activity is shown in the work by Thomas

Roth (2011) [7]. Roth has created a program that runs on Amazon's EC2 to brute-force wireless network passwords by testing approximately 400,000 passwords per second. According to his research, 'the average password is guessed in six minutes' at a cloud computing cost of $0.28 to $1.68 per minute. Previously, it would have been extortionately expensive to run such an attack, but based on the aforementioned examples, cloud computing makes these costs negligible [7]. These examples, demonstrate the potential of the computing power available as a potential launch pad for further attacks.

Similarly, the large volumes of data stored in the cloud make it a highly attractive ultimate target to attackers [4]. There are many references in the literature [4, 8] relating to the increasing number of globally reported cyber-attacks that aims at stealing businesses data. These come in the form of the increasing number of targets who have experienced difficulties in accessing data, suffered from identity fraud or been the victim of phishing scams. There have also been a number of high profile attacks targeting large industries, which can have detrimental effects even if they are dealt with quickly and efficiently [9]. These effects include loss of client confidence, misuse of company resources and loss of revenue. For instance, the largest DDoS attack in Norwegian history disrupted the websites and online payment systems of five banks, three airlines, two telecommunication firms and one insurance company. This attack was committed by an individual highlighting the computing power that can be leveraged by a single source [6].

Security Researchers Mary Landesman and Dave Monnier [10] have reported a 'meaningful increase' in attacks on cloud hosting providers [insert ref here]. These attacks commonly use set-up or hacked accounts to deploy command and control servers to conduct malicious activities. It is reported that 47% of phishing attacks were from exploited web hosts. This is because by updating the configuration of a single web host hundreds or even thousands of websites can be infected with phishing pages. Known attacks have used such approaches to compromise nearly 20,000 websites. This demonstrates that attackers are exploiting cloud computing to access the computing power required for larger scale attacks [10].

It has also been seen that attacks against the routers that control traffic and provide the backbone of the internet are growing in line with other cybersecurity issues. DDoS attacks, such as those against Cloudflare and Spamhaus, are increasingly abusing the Simple Network Management Protocol (SNMP). Researchers have found that in the month of May 2014, fourteen separate DDoS attacks made use of SNMP amplified reflection attacks [11]. This emphasises the rate at which attacks are evolving and highlights the importance of constantly evolving defence systems. Issues are further complicated by the various deployment models of the cloud and the responsibility of the various parties for security in each of these setups [12].

This paper aims to investigate the latest defensive systems proposed for use against DDoS attacks targeting the cloud model. In Sections I and II, the key areas of virtualisation and intrusion detection and the relevant security issues with each are examined. Section III presents a classification of intrusion detection in the cloud and highlights the main challenges facing their deployment. Section IV explains how countermeasures proposed for traditional networks are ineffective in cloud environments. Section V present the latest developments in the areas Virtual Machin (VM) security. Section VI, presents intrusion detection and prevention systems in cloud systems. Section VII focusses on defence systems against DDoS in the cloud. The security issues across each of these areas discussed in Sections V to VI, are investigated along with proposed solutions. A summary of the proposed solutions across these areas is presented in Section VII. Section VIII concludes the paper and highlights future research avenues.

II. Overview of Cloud Technologies

A. *Virtualisation*

Virtualisation is the key concept behind the cloud computing model. It allows programs to be portable across platforms as well as regulates their scalability, monitoring and security [13]. This technology provides interfaces, allowing for virtual process or system machines to be mapped onto the underlying hardware. Such interfaces allow guest instructions provided by the user of the Virtual Machine (VM) to be converted into host instructions through Dynamic Binary Translation. Instructions are converted in blocks rather than individually to provide greater efficiency and allow them to be saved for reuse in software caches [14].

A key benefit of virtualisation is that it allows multiple users to co-habit a single physical machine. This leads to resource consolidation, increased capacity, mobility and makes the system easier to maintain. Virtualisation efficiently supports many tenants, while attempting to isolate them from each other. It provides load balancing through dynamic provisioning and allows the migration of VM's between physical resources. Simultaneously, virtualisation 'poses a major security risk' given the difficulty in ensuring that different instances running on the same physical machine are fully isolated [15]. Vulnerabilities in the VM or VM Manager (VMM) can be exploited to bypass security restrictions or to gain unauthorised privileges.

Loganayagi et al. [16] argues that securing virtualisation technologies will improve cloud security. The key security mechanism of VM is isolation. Isolation allows multiple users to co-habit the same physical host without data leakage occurring across users and without unauthorised users being given access to the VMs of others. However, the scalability features of VMs can still allow some issues to be exploited, e.g., expose memory and process management functionalities leading to privilege escalation attacks. Defence approaches involving isolation can be split into those that isolate the running of VMs and those that focus on the isolation of shared resources [17].

B. *Cloud Architecture and Service Deployment Models*

According to Jang-Jaccard [4], the typical architecture of a cloud computing environment can be divided into 4 layers (see Figure 1):

- ***The hardware layer:*** This is usually a data centre with responsibility for the physical cloud resources including, servers, routers, switches, power, cooling systems, etc.
- ***The infrastructure layer:*** This layer is also known as the virtualisation layer and is used to pool resources by partitioning the physical resources. This is an essential component of the cloud computing architecture, as dynamic resource assignment and other key features are only made available through virtualisation.
- ***The platform layer:*** This layer consists of operating system and application frameworks, reducing the load of deploying applications directly into virtualisation containers.
- ***The application layer:*** This is the domain of applications, which can make use of the automatic scaling features of the cloud.

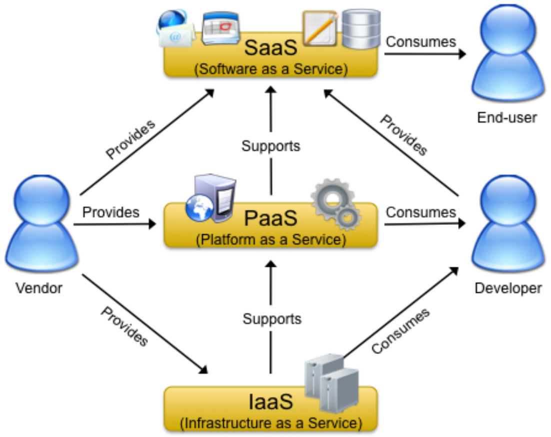


Fig. 1. Service deployment models of cloud computing [18]

There are many different deployment models offered by cloud providers to adapt to client requirements. These are typically divided into three groups [15]:

1) Infrastructure as a Service (IaaS) provides the client with access to hardware in a 'rented' capacity. The space on hardware dedicated to a particular client will vary depending on their demand, allowing the hardware to deal with fluctuations and spikes in requirements. Out of the box, IaaS provides only a basic level of security, e.g., perimeter firewall, and therefore applications deployed in this manner will need higher levels of security provided at the host. The responsibilities in this model vary between providers with some taking responsibility for the security of the infrastructure including the hypervisor and below, with the rest being left to the client.

2) Platform as a Service (PaaS) allows users access to virtual operating systems on top of the hardware layer. In such a setup, the CSP would typically be responsible for the security of the hardware and operating system elements. The client is responsible for the security of any further software that would be installed on top of these elements, i.e., at the application layer, the responsibility for security lies with the client. The provider may use metrics to measure the security of the applications deployed on their services. Below that in the hierarchy, it is up to the provider to offer strong assurances that data will not be accessible to other applications.

3) Software as a Service (SaaS) adds a further layer on top of the PaaS, with software being provided on top of the virtual operating systems. The security of the software is a priority as they provide a gateway to further down into the cloud structure. In this model, the client is dependent on the CSP for security measures, because company data is stored at their data farms. This storage could be located or replicated to aid availability to anywhere in the world.

Across all models, the security complexities are amplified due to the composite relations between the different deployment levels and user requirements. Security responsibilities needs to be particularly considered when using IaaS and PaaS for the development of other IT products to avoid introducing further security issues. Often, all deployment models and their specific responsibilities will be outlined in the QoS contracts drawn up between all relevant parties.

C. Cloud Vulnerabilities

The openness, elasticity, and amount of data stored in clouds make them attractive targets for attackers. The expansiveness of cloud operation across geographical and technological regions also brings with it the security issues associated with each of these areas. Traditionally, networks were less distributed, making defence mechanisms focused on insider attacks. Given that networks are now far more de-centralised and are connected globally via the Internet, the security risks in these systems have grown exponentially. Attackers are now able to target globally without concern over geographical location. Scripts and tools for launching attacks against networked systems are readily available on the Internet and require minimal user skills to execute.

IT users want to be assured that their data is secure. In the cloud computing paradigm, security responsibilities are passed to an outside agency. This can leave users feeling vulnerable, because they no longer control the physical storage where their data is residing and the legal frameworks around its protection. There is also the fear of limited availability or the introduction of further vulnerabilities, e.g., SQL injection and buffer overflow, which can be exploited through web browsers. Moreover, user interactions with the cloud are governed by traditional Internet protocols, e.g., HTTP, which makes it more difficult to identify attackers and easier for attackers to implement distributed attacks [19].

This paper pays particular attention to the vast increase in the number of DDoS as detailed by Wang et al. [8]. DDoS is an attack on system's availability to serve legitimate users. These attacks can be constructed in a number of ways. Table I summarises the main attacks aimed at causing this kind of disruption.

TABLE I. COMMON DDOS ATTACKS

Type of Attack	How it works
Flooding	Flooding can occur through all network and application layer protocols, e.g., HTTP, TCP, UDP, ICMP, etc. It attempts to saturate the network bandwidth by sending a large volume of packets from single or distributed sources so that it is unavailable to process legitimate user traffic. Flooding can be direct attack against the network or application, or reflective attacks via zombie machines.
Spoofing	This approach is used to falsify the origin of a network traffic to bypass filters, hide the source of an attack or gain access to restricted resources or services.
User to root	Aims to gain administrator (root) access privileges for a non-authorised account.
Port Scanning	Provides a list of open ports and the services provided by each; these can then be targeted by other attack methods. Port scanning is used in the first stages of an attack cycle and comes in many forms such as TCP SYN, TCP ACK, TCP ECHO, TCMP SWEEP, etc.
Oversized XML	The attacker sends a very large XML document (several megabytes in size) that contains elements, attributes or namespaces with large names or content. The Document Object Model parses documents into memory in their entirety to be analysed increasing memory requirements by a factor of 2-30.
Coercive Parsing	The attacker sends malformed XML aimed at clogging up CPU cycles by incorporating many namespace declarations or by simply using very deeply nested XML structures.
Web Service-addressing Spoofing	This is an extension of the spoofing attack, where the ReplyTo or FaultTo address in a SOAP header is falsified leading to a reflective attack.
Reflective attack	Request messages are sent to reflector machines via zombie machines containing the spoofed source IP address of the victim. The genuine replies to these requests are then sent to the victim causing flooding. Such attacks include ICMP ECHO reply flood, Smurf attack, Fraggle attack, DNS flood and SYN ACK(RST) flood.

In traditional networked systems, the disruptions caused by a DDoS attack against a particular target is limited to that target's resources or services. When this paradigm is transferred to the cloud, the potential for disruption crosses traditional organisational boundaries. This is because data is managed by a common CSP and the data of different organisations may be stored on the same physical hardware. Therefore, the scalability of cloud computing is what presents its main security challenges when compared to traditional networks.

III. INTRUSION DETECTION TECHNIQUES IN CLOUD

A. Intrusion Detection

Intrusion detection is the first step in identifying a malicious behaviour against a system. The key challenge is to reliably differentiate between legitimate users and attack traffic. There are two standard approaches used by Intrusion Detection Systems (IDS), these are knowledge-based intrusion detection and behaviour-based intrusion detection [20].

Knowledge-based systems must possess an attack description, typically a signature that can be matched to attack manifestations. Signatures range from simple pattern matching to network packets, such as those used in BRO, ASAX and NADIR systems [21], building up to neural networks that map multiple sensor outputs to abstract attack representations. Jang-Jaccard [4] argues that signature-based systems are ineffective given the constantly evolving landscape of cyber-attacks. While it is true that this approach requires constant updates as new signatures are identified, the simplicity of its structure allows it to be rapidly deployed across systems. In addition, the knowledge databases used provide effective attack classification allowing a more directed response to be triggered.

Behaviour-based, also known as anomaly-based, systems are designed to evolve to meet new, previously unseen threats. They involve monitoring network attributes and assume that the behaviour of malicious parties is noticeably different to that of legitimate users. This assumption allows anomalies to be flagged and alarms to be raised. These systems often require a training period to build a model of network attributes, which raises the cost of their implementation. The usability of these systems is dependent on the False Alarm Rate (FAR) that they generate, which is made up of both false-positive alarms (raising an alarm for legitimate traffic) and false-negatives (the failure to register an intrusion attacks). This approach can struggle to classify attacks allowing only a general system response to be triggered, yet, it can respond to previously unseen cyber-attacks.

Compared to IDS, Intrusion Detection and Prevention Systems (IDPS) include preventative measures to stop attacks in real time rather than simply detecting them once they occur. These systems follow the design principles of IDS, but also take preventative action such as logging a user off, initiating system shutdown, halting the system or disabling connections [22].

B. IDSs Classification and Challenges

In general, IDS systems designed for cloud computing can be classified into four main categories [23]:

1) Host-based IDS (HIDS): These systems monitor and analyse log files, security access and user login information to detect intrusive behaviour.

2) Network-based IDS (NIDS): These systems monitor IP and transport layer headers with behaviour being compared with previously observed behaviour in real time. This approach does not work with encrypted network traffic.

3) Hypervisor-based IDS (HyIDS): These systems allow users to monitor and analyse communication between VMs, within the VMM-based virtual network and between the VMM and VMs. These systems benefit from an availability of information that can be analysed to detect intrusions.

4) Distributed IDS (DIDS): DIDS consists of a number of IDSs (HIDS and NIDS) placed across a large network. These individual IDSs communicate with each other via a central analyser, which aggregates system information from the different IDSs. This system benefits from the qualities of both HIDS and NIDS to detect known and unknown attacks. However, there is a high computational cost in the communication between these systems. In a cloud computing *environment, the central analyser can be placed on a host machine or at the processing server. However, if the analyser is compromised or unable to communicate, then the system will not be able to react to further threats.*

IDPS can prove to be an invaluable tool in the early detection of malicious activity helping to prevent attacks from succeeding. They can also gather forensic evidence. However, traditional IDPSs are largely inefficient when applied to cloud computing given its openness. Patel [12] investigates the requirements of IDPS in the cloud architecture given the ineffectiveness of traditional methods by asking what criteria and requirements should an IDPS meet to be deployed on the cloud? Which methods or techniques can satisfy these requirements? The list below outlines some of the challenges that traditional IDPS struggle to counter:

- They do not scale to deal with cloud requirements and do not satisfy the requirements of high-speed networks.
- The traffic profiles of networks changes frequently rendering the audit data used to train the IDPS unsuitable very quickly.
- They generate high false alarm rate [24].
- There is no uniform standard or metric for evaluating an IDPS, which can often lead to misleading information as to their effectiveness.
- It is very difficult to identify internal intrusion attacks given that correctly configuring the systems and implementing organisational policies is a difficult task.

IV. Distributed Denial of Service (DDoS) Attacks and Traditional Countermeasures

The most common attack vector that has been used to attempt to adversely affect cloud services is DDoS attacks. A DDoS attack aims to render the computing resources of the victim unavailable by modifying the system configuration or by sending it too high workload. Moore [25] determined that in 2006 the average rate during a DDoS attack was 500 requests per second and that attack typically lasts less than five minutes. These figures may well now be considered out of date with Goel et al. [26] noting attacks with a rate of greater than 100Gbs.

DDoS attacks can be divided into two general categories, application level attacks and infrastructure level attacks. In application level attacks, e.g., HTTP flood, zombie machines establish TCP connections to the victim server and send legitimate requests. Systems used to detect such attacks can struggle to differentiate between attacks and busy periods, such as the start of the workday where many legitimate users may attempt to access resources simultaneously. One method for dealing with these attacks is the use of CAPTCHA puzzles. However, these puzzles are only suitable for an initial user login or registration; the user may become frustrated with the service if they are used any more frequently.

Infrastructure level attacks require the attacker to send a flood of packets to the victim server in to saturate or bottleneck the victim so that it can not respond to legitimate requests. For this type of attack only the victim’s IP address needs to be known. Typical direct infrastructure layer attacks include TCP flood, UDP flood, ICMP flood and SYN flood.

Traditionally, well-known countermeasures have focussed on dealing with DDoS attacks through a variety of methods devised around the questions [27]: (1) Where is the attack detected? (2) How is the attack detected? (3) What is the response mechanism? (4) Where to apply the response mechanism? (5) Where is the control (decision) centre from which filtering rules are taken?

The most common DDoS defence approaches combine elements located in the source-end and victim-end in to combine their advantages. However, the use of multiple components leads to gaps in coverage, which can be exploited. The source-end is the location from which the attack is launched; this is the best place to intercept an attack as it causes the least disruption to legitimate traffic. However, distinguishing between legitimate and malicious traffic at this point is a serious challenge. D-WARD [27] is a system that employs a firewall at the source end. It gathers 2-way statistics from the border routers. This introduces significant overhead because D-WARD is continuously monitoring and classifying traffic based on IP address, comparing statistics and applying filtering rules. The operation of D-WARD affects the speed of the entire network whether there is an attack or not. Beitollahi et al [27] suggest that there is no benefit for deploying source-end firewalls considering the overhead and performance loss they introduce. Yet, a user would not want his networks to be compromised and turned into a pad to launch further attacks. This is more critical when considered in the cloud computing paradigm, where undetected intrusion has the possibility of giving an attacker access to a far greater amount of resources than a traditional network could provide. The CSP would need to balance the the threat of becoming a source of an attack with the detriment in service provided to legitimate users.

DDoS traffic is easier to identify at victim-end points. IDPS at these points are effective at generating attack signatures, which they can then be used by upstream routers to rate-limit or filter traffic. However, by this point the bandwidth of the network could be saturated. Moreover, these infrastructural approaches require the cooperation of multiple Internet Service Providers (ISP) to cover the required range of administrative domains. There are also issues with security and authenticating the communication channels.

For the reasons given above, the majority of traditional DDoS countermeasures are ineffective against application layer attacks. This is mainly because the packets have been transferred and the TCP handshakes have been completed meaning that the packets appear to be legitimate. Packet sniffing protocols are therefore ineffective at this level.

V. VM Common Security Vulnerabilities and Defence Mechanisms

Virtualisation is the key underlying technology of the cloud computing model and therefore its security needs to be considered as the foundations of any proposed system. Particularly, the fundamental weaknesses in the VM architecture need to be addressed to enhance security across other layers. For instance, attackers on the same physical machine can use malicious code to get control of other VMs. They can then deploy a class of rootkits, e.g., UMBR, which operate under the operating system. These rootkits cannot be removed easily and are difficult to detect [13]. There are also attacks targeting VM migration.

Isolation is the key security feature to protect VMs from malicious attacks. The isolation-based defence approaches can be split into those that isolate the running of VMs and those that focus on the isolation of shared resources [17]. The first approach can limit the ability of the system to schedule the work of legitimate VMs. To implement the second approach, a monitoring and mediation mechanism is needed to probe all resource requests and to allocate these requests to VMs. In addition, to implement the strict policies needed to enforce isolation takes many OS hooks and is difficult to enforce across a large-scale distributed system.

Volokyta [28] suggests a VM Monitor to secure VMs. The proposed system intercepts system calls and maintains a log file of system warnings. The authors do not give the practical design details and experimental results. Therefore, it is not possible to comment on the performance of this system. However, VM management systems are frequently part of security design specifications.

Yu et al. [17] consider Chinese Wall properties, where an object can be read if the subject has accessed a prior object from the same dataset or the objects conflict of interest is set to new. This system records the behaviour of VMs to obtain traces used to calculate the Aggressive Conflict of Interest Relation (ACIR) or Aggressive in Ally with Relation (AIAR). Isolation rules are combined with constraint relations to get the access matrix, which records the maps to give dynamic updates between the VMs and hosts. An algorithm is implemented to guarantee isolation between conflicting users. In this approach, specific monitoring systems are not required, but a trade-off is made between security and resource utilisations. To further improve this system, more efficient methods for conflict analysis are needed. A suggestion for enhanced isolation using ACIR and AIAR to describe constraint relations has been put forward by in [17].

Lui et al. [29] suggest a framework for enhancing VMs security in clouds using module measure. Metrics of the executables running in VMs are taken and compared to a reference table of trusted measurements. This aims at combating user-level security in SaaS, where many individual users access a single instance of an application. In this framework, a trusted VM is used to monitor other VM instances, meaning that the status of the measurement module needs to be noted to ensure that the system can be trusted.

Another approach, presented by Williams et al. [30] is to use N-version programming in the construction of VMs. The authors introduce diversity in VM design to avoid a sequence of events that leads to failure. This approach lends itself to automation making it scalable, but can make the development of compatible systems difficult. The proposed structure provides diversity during execution through Address Space Randomisation (ASR). This approach does not remove vulnerabilities, but aims to make them more difficult to exploit, because an attack that works against one binary will not work against another. This means that only a single instance of an application will be affected during each attack.

To recapitulate, virtualisation poses a number of security issues that need to be addressed. A benefit of virtualisation is its ability to allow users to isolate VMs and resources, which, theoretically, enhance security. A number of solutions have been proposed to monitor and enforce the principles of isolation and thus secure the virtualisation layer. A common suggestion is to use a VM as a designated manager to monitor the operation of the other VMs in the network. If only a single management machine is used, then this could create a bottleneck in the system especially in an architecture where additional VMs can be generated and deployed autonomously. The systems reviewed could be made more suitable for a cloud environment if monitoring VMs could also make use of the scalable nature of the cloud. This would mean that they could increase their number autonomously to ensure efficient management of resources and monitoring of isolation throughout periods of operation.

VI. IDPSS IN THE CLOUD

IDS and IDPS face difficulties when transferred from traditional networks to cloud-based designs. Issues such as those with their deployment locations and the separation of legitimate traffic from malicious traffic pose real challenges with their implementation. Defence systems must successfully determine the need for their use before they can be executed, otherwise the user experience will deteriorate. The latest developments in IDPSs are investigated in this section''.

Al-Jarrah et al. [31] suggest embedding the temporal behaviour of attacks into a Time Delay Neural Network (TDNN) model to defend against probe or reconnaissance attacks. The suggested system works on a universal IP plan as the relationships between the inputs are the keys, and no range or class of IP addresses is used. This makes this system suitable for use in cloud computing given the scalability that it offers. However, an autonomous method for including relationships for newly generated nodes would need to be created. The experimental results show the approach to be effective when tested against the DARPA Intrusion Detection Evaluation [32] and other IDS systems such as SNORT. The overheads of the system are not discussed or compared to other techniques, though the authors state that their 'system is characterised by high throughput because after the system is trained, it takes constant time to detect any attack'.

Alqahtani et al. [22] put forward a system to prevent SQL injection in cloud computing web-based systems using signature-based approaches. They focus on the application layer, because web software services contain the majority of security vulnerabilities in cloud systems. Automated tools, such as SQLmap, are identified as having the potential to be used maliciously by hackers to attack cloud-hosted databases.

The presented evaluation method provides suitable metrics for measuring the quality of an IDPS system. These include vulnerability detection, average response time and number of false positives. An improvement of this system is to implement these measurements in other test designs to provide suitable data for comparing IDPS systems.

SNORT and OpenFlow are combined by Xing et al. [33] to produce an IDPS (SnortFlow) that can reconfigure a cloud network system in real time using Iptables. The IDPS is made up of four components:

1) Cloud cluster – this is based on the efficient parallel virtualisation solution, XenServer.

2) Open Flow Switch – this connects resources on different cloud servers.

3) Open vSwitch – this is the software version of the switch and is implemented in one of the domains of the Xen hypervisor.

4) Controller – this provides centralised control over the enabled infrastructure. A POX controller is easy to program and can synchronise both physical and virtual networks.

This approach considers the status of the network when deciding, which actions to take based on the findings of traditional IDS approaches. The optional response is selected using the graphical attack IPS NICE [33]. This decides if a response is necessary and chooses from options including, traffic redirection, traffic isolation, deep packet inspection, MAC address change, IP address change, block port or quarantine. However, every packet is monitored by SNORT, which could create a bottleneck, especially under large-scale DDoS attack conditions. The packet dropping rate of SNORT increases dramatically once the traffic rate exceeds 45,000 packets per second [34] thought it does have a better throughput than some other similar systems such as Suricata [35].

Hassani [36] combine IDS, IPS and hybrid detection techniques (pattern matching and anomaly detection) to address the issues of each individual approach. This effort focuses on distributed attacks coming through the infrastructure layer. Individual system components each have their own weaknesses, which may result in some attacks going undetected if attackers alter the timing used during attacks to make these appear as random individual requests. The collaborative IDS put forward makes use of the entire network to correlate events and to deduce a distributed attack that occurs in several places. There is no implementation of the suggested system, which makes it is difficult to measure the efficiency and cost of this approach.

Most current Intrusion Response Systems (IRS) use static matching to decide a suitable response action to an attack. The issue with this approach is that it does not consider the status of the entire system. Alazab et al. [37] suggest a system to improve detection efficiency. The proposed system uses an Intelligent IDS (IIDS) built up of SIDS, AIDS and IRS. The mechanism links the state of an attack with a response to raise an alarm, or to audit, hold, abort, disconnect or refuse packets. The IRS uses two stages to assess the potential risks of the anomaly using the Microsoft DREAD model shown in Table II.

Initially, the SIDS examines the content of the user request for currently known intrusions. This is followed by the AIDS step to accommodate the shortfalls of the SIDS. The AIDS assumes that any request received from the user is an anomaly unless proven otherwise.

TABLE II. MICROSOFT DREAD MODEL [36]

Damage Potential	How great is the damage if the vulnerability is exploited?
Reproducibility	How easy is it to reproduce an attack?
Exploitability	How easy is it to launch an attack?
Affected Users	As a rough percentage, how many users are affected?
Discoverability	How easy is it to find the vulnerability?

A risk assessment matrix is then used to determine whether the request is fulfilled or which action is taken. The IIPS is flexible enough to accommodate different web application architectures. This is aided by the fact that the communication of the SIDS and AIDS are based on web application architecture.

VMFence is a system put forward by Jin et al. [34], which uses a VM Monitor based IPS to monitor network flow and file integrity in real-time. The defence of the network and file integrity protection varies with the state of the VM. This approach is based on the fact that virtualisation-based cloud computing comes down to the security of virtualisation itself. The system uses a privileged VM and contains 5 phases:

1) Detection – This captures all network packets and dispatches them to other detection processes according to their MAC address.

2) Policy Updating Component – Used for intrusion response and collects all alerts.

3) Front-end to Back-end Communication – Updates the firewall rules in real-time.

4) File Integrity Monitoring – Observes read/write operations.

5) Notification – Receives service type and sensitive files defined by cloud users. This unit also collects alerts for the cloud provider.

Snort is used as an IDS. Iptables are used as a firewall with policies being updated via a shared page located in the XenStore. The back-end and front-end communicate via the event channel. This approach is faster than traditional response by network. File integrity is monitored on the blktap mechanism that can directly manage disk activities with small performance overhead. VMFence uses a privileged VM to monitor the other VM nodes, which cause a bottleneck in the system. However, this system can make use of scalability properties of the cloud to relieve any bottlenecking.

The reviewed solutions in this section outline many issues with the design of current IDS and IDPS systems. Common themes can be drawn across all of the proposed systems regarding the size of system overheads, how to react once an alert is triggered and how to effectively reduce the false alarm rate. Many proposed systems focus on detecting a single style of attack or protecting a single layer of the cloud architecture. Although these approaches can be successful in providing security to one part of the cloud, a cohesive and adaptable system is required to avoid the layering of individual security components. The use of many individual defence systems can lead to the development of further vulnerabilities along the protection vulnerabilities of each of these systems. There is no standardised procedure for measuring the quality and effectiveness of an IDS making comparisons between proposals difficult. Implementing a standardised set of metrics would enable performance comparison of various systems and allow the strengths and weaknesses of each to be identified as well as highlighting areas for further research in the field.

VII. Systems For Defending Against Ddos in the Cloud

DDoS attacks can render CSPs unable to provide their users with the service as outlined in their QoS documents and/or they have their resources manipulated to launch an attack against external targets. This section builds on the knowledge of the cloud architecture and IDPSs to analyse proposed systems aimed at protecting the cloud against DDoS attacks.

Yang et al. [38] propose a trace-back and filter system to protect the cloud from DDoS attacks. The current packet tracing methods of Probabilistic Packet Marking (PPM) and Deterministic Packet Marking (DPM) will become ineffective with the introduction of IPv6. To overcome this, trace-back is implemented by adding a tag to Service Orientated Architectures (SOA) packets to record the route taken. This method is rather limited in its effect to counter DDoS attacks, because the tag is only added to the packet once it is relatively close to the server. When an attack is launched via the Internet, SOA packet tracking does not provide enough information to identify the source of the attack. The tests presented by the authors do not consider spoofed IP addresses or the use of zombie machines in the attack. The attacker does not need to cover the locations of zombie or spoofed machines, as the overall source of the attack will remain protected. Another weakness is assuming that even whilst under attack, the system will operate properly and communicate correctly with the upstream filters. These filters are expected to remove attack traffic, but the authors assumes that attack traffic is all coming from single sources. If the bandwidth is flooded, then the deployment of these resources can not be relied upon.

Another trace-back model is suggested by Joshi et al. [39] using DPM and training data to inform the filters in a neural-network. The future of DPM may be limited with the introduction of IPv6 and the fact that trace-back tags can only be introduced once a packet is within the cloud network. This system has a success rate of correctly identifying approximately 75% of attack traffic, though its ability to detect currently unknown attacks is not researched. It has a significant time variation in the detection rate of attack traffic from 20ms to 1s; an overhead that may cause disruption to legitimate users when accessing systems even if an attack is not taking place.

Karnwal et al. [19] introduce a filtering tree to act as a service broker within a SOA. They investigate the vulnerabilities in standardised cloud APIs and how they can be exploited when used in provisioning, management and monitoring of services. They propose adding a signature reference element to each SOAP request to ensure that it comes from a legitimate source. Double signatures are created using hashed characteristics of each SOAP envelope, such as the number of child or header elements. The client IP address is also maintained in the message header along with a puzzle that is stored as part of the WSDL file. Scanning each packet individually will eventually lead to bottlenecks. IP trace-back is put forward as a method for discovering the source of attacks and updating defence systems to drop packets from that location. However, the system will remain ineffective in dealing with flooding attacks from distributed sources or those using spoofed IP addresses.

Vissers et al. [40] work aims to deal with DDoS attacks at the application layer. Their solution aims to protect primarily against HTTP flooding, Oversized XML, Coercive parsing, Oversized Encryption and Web Service-addressing spoofing. A reverse proxy is added as a filter to intercept all service requests. It is claimed that this filter adds no overhead to the cloud operation and that users experience no effect to their service. The web service itself is only accept requests that come from the defence server. If the server itself is directly flooded, then the server’s reaction to legitimate requests will be affected. Initially, the system processes the HTTP header to get the size of the request, to see if the packet is oversized. Reading the header also means that the number of requests from a single client in a given period can be limited to enable fast detection of HTTP floods. To provide further security, strong authentication is imposed on users attempting to connect to web services. This helps protect against ‘meek’ attacks, where a large number of zombie machines make a low rate of individual requests that can collectively lead to DDoS. The SOAP action header is then taken and used to determine the requested operation of the packet without having to examine all of the XML content. The action header could be spoofed by malicious parties and therefore further checks are required to verify if this is the case. The second phase involves processing the key properties of the XML content and comparing them against the pre-determined attack models. WS-addressing spoofing is also extracted before the header details and content processes are compared to determine if they match. The results presented are encouraging, but are limited in the sense that only a single platform set-up is tested with a single-target web application. In addition, the attack tool could not generate attacks involving encryption or signatures and oversized encryption attacks were sent directly from legitimate sources. This system requires further development to cover multiple platforms and applications to reflect its performance in real world scenarios.

Another application layer DDoS IDPS specifically designed to deal with Low and Slow (LOS) attacks is presented in [41]. These attacks are rarely detected using pattern matching or threshold measuring techniques given their low resource consumption approach. The authors propose a reference-based architecture to mitigate DDoS attacks by utilizing a Software Defined Infrastructure (SDI). Senthilmahesh et al. [42], Mathew et al. [43] and Tang et al. [44] describe techniques used for detecting LOS DDoS in the proposed system. The system introduces a ‘healing’ approach that is implemented if an intrusion is detected. This approach involves migrating legitimate users from compromised to newly generated VMs. ‘Shark Tanks’ are introduced as quarantine areas for potentially malicious traffic. This allows suspect users to be monitored more closely, while continuing to receive a suitable level of application access in case a legitimate user is wrongly redirected. The locations of the Shark Tanks are disguised using OpenFlow switches, which can rewrite packet headers so that attackers are unaware that they are being monitored. The system and clusters are described using Domain Specific Language (DSL) and the use of VMs means that the system is scalable. Two concrete implementations of the design are put forward. However, the results of these implementations are not compared against existing systems or each other.

Wang et al. [8] use the autonomous generation properties of the cloud to base service delivery on randomly generated and assigned proxy nodes. This restricts the information that attackers can determine through reconnaissance attacks with all internal IP addresses being kept hidden. Attacking machines using spoofed IP addresses become ineffective, as they will not receive server reassignment messages. Strong authentication techniques are proposed to prevent external attackers from accessing proxy nodes. The application server will only accept requests from a designated ring of proxy nodes and from clients who have made a successful connection. Connections are monitored by tokens passed between the user and the authentication server. The presented solution focuses on preventing ‘insider’ attacks by imposing strong authentication to connect to a proxy node, which will stop external attackers from launching an attack unaided. The benefits of the proposed system are that it make use of the properties of the cloud allowing it to scale and it does not require universal deployment to provide protection. However, the system relies entirely on the IP addresses of key components, mainly the application server, being kept hidden but there is no explanation as to how this can be achieved. Currently, all of the datasets used to generate the models used by the system are stored centrally by the application server. The authors acknowledge as being an issue, the application server itself may become even more of a target. An effective fix to this problem is to distribute these datasets.

Wang at al. [8] propose a greedy algorithm to deliver a ‘near-optimal’ method for assigning users to proxy nodes. Users are ‘shuffled’ between newly generated nodes when an attack is detected against a proxy node. The repeated shuffling of users allows the system to identify the insiders provoking the attack. This work is extended by Jia et al. [45], who introduce a selection of algorithms to optimise runtime reassignment plans. They optimise the greedy algorithm in the form of a dynamic programming algorithm. The real-world implementation of this algorithm is very computationally expensive. Both solutions use the number of persistent bots, containing the intelligence to follow migrating servers, as a key parameter for calculating the optimised ‘shuffling’ pattern. However, in the real-world this value is unknown and can only be estimated.

Jia [45] extends the system proposed in [8] to provide security at the application layer and to offer security for anonymous users by removing the need for strong client authentication. This produces a generic DDoS protection product that can be deployed by non-ISP organisations. This product is efficient at mitigating DDoS attacks and is more cost-effective than static based systems. Both [8] and [45] do not discuss how the attack is detected by the proxy node, they only give the response that is used to identify and dispel the source. Implementation overheads are discussed and are dependent on the number of shuffles required. They are also dependent on the size of the geographical area covered, which could be global in a cloud computing context.

Another attempt at using the capabilities of the cloud in a DDoS defence is proposed in [17]. This system aims to protect individual cloud users by creating clones of virtual IPS to filter traffic. A queueing algorithm is defined to determine the number of IPS clones necessary to defeat the DDoS attack and maintain acceptable QoS. This system is based on the assumption that to defeat DDoS attacks, the defence system must have access to greater resources than those of the attackers. Compared to the research in [25], this is feasible when applied by a CSP. This means that DDoS attacks are currently unlikely to be able to affect an entire cloud service. However, the cloud resources available to individual clients are likely to be more limited making them still susceptible to these attacks. The authors assume that the number of service requests follows a Poisson distribution during both normal usage and attack periods, and that the rate of legitimate requests remains constant during both periods. Therefore, the average time that a packet is in the system provides a suitable measure of QoS. The theoretical results provided assume that the IPS cloning solution is effective and that the cloud contains enough idle resources to overcome the attack. The economic cost of implementation is also considered using the Amazon cloud (EC2) pricing model.

Huang [46] proposed a low reflection ratio mitigation system to be deployed in front of the IaaS. The system consists of Source Checking, Counting, Attack Detection, Turing test and Question Generation modules. In the implementation of their defence system, the authors take into account the challenges of computational efficiency and overheads and their effect on legitimate users. The Turing test is embedded in the kernel and uses text-based questions generated using Lexical Function Grammar (LGF). This approach requires less bandwidth than the more traditional image-based puzzles, such as CAPTCHA. A blacklist, whitelist, block list and unknown are used to categorise incoming packets based on IP addresses. These lists are maintained by administrators through APIs. The use of these APIs opens the system to malicious manipulation from insiders. Although this system uses the cloud capabilities to provide protection against bottlenecks, it incurs an operational degradation of 8.5% when monitoring traffic against a blacklist of 100000 addresses.

Fujinoki et al. [47] build on the limitations of overlay networks in hiding the location of target servers through the use of gateway routers. The suggested Dynamic Binary User Splits (DBUS) system protects clouds from insider attacks and compromised user host machines. DBUS avoids the need for migration of network items to other hosts. It also removes the need to monitor all network traffic, which provides lower computational overhead. Each proxy router contains a Bloom filter, which is a data structure that can efficiently test for the presence of certain values. A user management table is used to hold records of which users are assigned to which proxy nodes and no user can return to a previous proxy once they have migrated. When an attack warning is issued, more user proxy machines are deployed with the number of users assigned to each being halved until attackers are identified. Simulations results proves the DBUS a promising system. Yet, there is a need for this system to be implemented in a real-world environment and evaluated under real-world conditions.

Tripathi et al. [48] investigates the use of the open-source tool Hadoop in providing a DDoS defence. Hadoop provides tools that use the MapReduce framework for processing large amounts of data in association with the Hadoop Distributed File System (HDFS). The authors approach overrides the traditional First in First Out (FIFO) scheduling mechanism with a Self-Adaptive MapReduce (SAMR) scheduling algorithm, which divides jobs into tasks that are then assigned to map nodes. The benefit of SAMR is that it also reads the historical information that is stored on each node and adjusts its task distribution based on this information. By measuring nodes performance, the task execution time can be improved by up to 25%. Hadoop is capable of efficient network behaviour analysis. However, its current implementation needs further optimisation to be used for cloud defence.

An alternative method for detecting DDoS flooding attacks is presented in [49]. A distance estimation technique is used to estimate traffic rates. The distance value is calculated using the Time-To-Live (TTL) of a packet. The majority of operating systems only accept initial TTL values of 30, 32, 60, 64, 126 and 255 making the estimated distance the smallest of these values that is greater than the current TTL value. Exponential smoothing is then implemented to provide the real-time measurement of the roundtrip of IP traffic. Finally, absolute deviation is used to determine if the behaviour is abnormal. This approach attempts to avoid the reliance on attribute dependences that can be spoofed or the time delays associated with traffic monitoring. The authors suggest that ISPs should be responsible for implementing filters on traffic as they receive the packets. As previously discussed, this practice is unlikely to be adopted.

Latif et al. [50] review approaches to protect against DDoS attacks focussing on Wireless Body Area Networks (WBANs). WBANs devices are limited in computational power, available bandwidth, security and battery life. This makes them ideally partnered with the cloud, where much of the complex computational requirements can be moved. Due to resource restrictions, there is a need for minimal overheads in any WBAN DDoS defence system. A number of approaches has been proposed to address this requirement. For example, the system described in [51] places IDS' at different locations in the cloud space and then has them collaborate to share attack alerts. This approach assumes that a node will have the available bandwidth to send an alert when it is under attack. In [52], a simple IDPS that uses a statistical method to create and apply a covariance matrix of network behaviour is proposed. Current network behaviour is compared to the created model, while the TTL of packets is used to identify the source of the attack. In [53], a similar behaviour-based system featuring a training period is presented. The system computes a score for each packet. These scores are then used to determine which packets to drop in an attack scenario. This approach delivers a high-speed system with minimal memory requirements and an acceptable level of filtering accuracy. This makes it suitable for real-time implementation. A correlation pattern detection module was added to this system by Priyanka et al. [54] to overcome the flaws in the Confidence-based Filtering (CBF) by introducing a confidence value to the packet header.

As shown in this review of the recent IDPSs, there have been many suggestions for tackling DDoS attacks against the cloud computing paradigm. The cloud architecture pose many security vulnerabilities at different level, which resulted in solutions being proposed to primarily defence against a single type of or point of attack. An important step forward is the utilisation of the cloud capabilities in the design of defence systems. This enabled systems to adopt the scalability properties of the cloud to enhance the security for all parties. It is important for security solutions to provide protection for individual clients and their services as well as the cloud as a whole. Commonly, individual client applications and web services will be the targets of DDoS attacks. Individual clients will only notice issues with their own QoS and these are the issues that will further perpetuate the security fears of adopting cloud computing architectures. To develop a comprehensive defence system, aspects of these research solutions need to be integrated in one product to protect against a wider range of attacks. System designers should pay particular attention to the 'secure' integration of the cloud underlying technologies to avoid introducing further vulnerabilities to the cloud architecture.

VIII. Summary of Features – DDos Cloud Protection Systems

In this section, we present a summary of all man concepts discussed in previous sections. The reviewed different attacks are listed with their corresponding response mechanisms. The recent solutions for various security issues are also grouped into logical categories to make gaps in the literature more obvious. This is followed by a taxonomy that attempts to classify DDoS protection systems, which have been proposed for the cloud computing paradigm. A description of each classification category and the order in which they have been applied is given in Table IV.

Figure 2, translates the taxonomy given in Table IV to a flow chart showing the exiting DDoS cloud protection systems and comparing the implementation of different features in the proposed systems. This allows us to see common techniques and highly secure facets of the systems, while also highlighting weaknesses and the areas of focus for future work in these areas.

TABLE III. SUMMARY OF CLOUD PROTECTION SYSTEMS KEY FEATURES

<table>
<tr><td colspan="2"><i>A.</i> <b>Intrusion Detection</b>
Description: All systems identified in this survey make use of intrusion detection to raise an alarm and to determine which response to adopt. Currently, statistical models are the most widely used methods for detecting intrusions. Other methods include the use of Artificial Intelligence and Neural Networks.</td></tr>
<tr><td><b>Knowledge-based IDS</b></td><td>• Used for identifying previously known attacks.
• Attack signatures stored in databases.
• Simple to implement.
• Require frequent updates to maintain security level.
• Slow to react to new attacks, as new behaviours signatures need to be added to the relevant databases.
• The aim is to define suitably abstract signatures that can potentially recognise new attack designs based on previously identified patterns.
• Provides effective attack classification, which allows a more targeted response to be triggered.</td></tr>
<tr><td><b>Anomaly-based IDS</b></td><td>• Monitors network attributes.
• Assumes that malicious network behaviour is noticeably different to regular behaviour.
• Able to detect unknown attacks.
• The usability of these systems is dependent on the false alarm rate.
• Requires a system-training period. The training period needs to be carefully selected to represent standard network behaviours.
• Models can be updated with new information while deployed. This may be required as business practices evolve.
• Data analysis tools, such as Hadoop, can be used to create models and monitor real-time behaviour [48].
• Suitable behaviour models are difficult to create given the user flexibility that the cloud introduces.
• Can struggle to classify the type of attack meaning that only generalised responses can be issued.
• Greater implementation complexity.</td></tr>
<tr><td><b>Hybrid</b></td><td>• A combination of knowledge-based and anomaly-based IDSs used to combine the strengths of both of types of approach.
• Highest level of complexity to implement.
• Higher overheads as packets are monitored through multiple types of IDS.</td></tr>
<tr><td><b>Deployment</b></td><td>• Deploying IDSs close the server makes identification of attacks easier. It also limits the effectiveness of countermeasures.
• Deploying IDS further from the server or external to a perimeter firewall makes the identification of an attack more complex. However, this allows countermeasures to have a greater impact.</td></tr>
<tr><td colspan="2"><i>B.</i> <b>Responses</b>
Description: The reaction of a system once an intrusion or attack has been detected. This can involve responses to neutralise an individual attack and the introduction of preventative measures to secure the system against future attacks of the same nature. These responses have been summarised into categories.</td></tr>
<tr><td><b>Filtering</b></td><td>• Update upstream filters to block traffic once the source is identified. This assumes that there is enough bandwidth to send these messages. This may be compromised in a DDoS attack scenario.
• Deployment locations can greatly affect response outcome.
• Update firewall protocols to include new responses.
• Add information to packet headers to identify legitimate packets [19].</td></tr>
<tr><td><b>Rate Limiting</b></td><td>• Attempts to relieve the pressures on bottlenecks.
• This affects all network traffic, not just the malicious.</td></tr>
<tr><td><b>Adapt use of Virtual Machines</b></td><td>• Using cloud features (scalability) to increase/decrease number of VMs as required [17].
• Increase the number of VMs to enhance isolation (DBUS - [47]).
• Logical and physical migration of resources [8, 45].</td></tr>
<tr><td><b>Identify attack source</b></td><td>• Use trace-back techniques to identify the source of attacks [39].
• Trace-back systems have limited effectiveness against multisource distributed attacks
• Can fail to identify ‘spoofed’ IP addresses.</td></tr>
</table>

C. **Management** Description: How are different aspects of the proposed system managed to ensure security.	
Authentication	• Strong authentication for legitimate users. Using image-based (CAPTCHA) or text-based puzzles [46]. • Authentication needs to avoid being obstructive to the user experience. For example, puzzles are suitable for logins but should not be used for routine tasks.
System Monitoring	• Several systems use VMMs to monitor the state of deployed VMs. Having these dedicated machines is a useful consideration but they risk becoming a bottleneck in high traffic situations [34]. • Using the scalability features of the cloud to deploy further VMMs as required can reduce the risk of bottlenecks [8].
Overheads	• The majority of systems aim to scan each packet as it enters the cloud network. This can introduce large overheads into the system affecting the users QoS. If only a single machine is allocated the task of packet monitoring, this can create a bottleneck in the system. • To reduce detection overheads, many systems aim to detect a limited set of attack traits.

TABLE IV. DDOS CLOUD PROTECTION SYSTEM TAXONOMY

<table>
<tr><th colspan="2">Taxonomy Layers</th></tr>
<tr><td><b>IDPS</b></td><td>There are two fundamental approaches to intrusion detection used by the protection system. These are typically either knowledge-based that use signatures to recognise attacks that have previously occurred, or anomaly-based that make use of data models to identify suspicious network behaviour. Anomaly-based systems have the advantage of being able to identify previously unseen attacks; however, they can suffer with high false positive rates.
<table>
<tr><th>Knowledge-based Intrusion Detection</th><th>Anomaly-based Intrusion Detection</th></tr>
<tr><td>• [19] Karnwal (2012)
• [46] Huang (2013)
• [47] Fujinoki (2013)
• [17] Yu (2014)
• [48] Tripathi (2013)
• [54] Priyanka (2013)</td><td>• [39] Joshi (2012)
• [40] Vissers (2014)
• [41] Shtern (2014)
• [8] Wang (2014)
• [45] Jia (2014)
• [49] Chopade (2013)
• [52] Ismail (2012)
• [53] Chen (2011)</td></tr>
</table>
</td></tr>
<tr><td><b>VM Management Point</b></td><td>Compares the use of VMM modules to manage systems. Classification is based on whether a single VMM is used to monitor VMs regardless of the scale of the cloud resources being used, or whether the number of VMMs is increased when required based on cloud scaling principles.
<table>
<tr><th>Scalable VMM system used (Distributed)</th><th>Single VMM used (Centralised)</th></tr>
<tr><td>• [17] Yu (2014)
• [41] Shtern (2014)
• [8] Wang (2014)
• [45] Jia (2014)</td><td>• [39] Joshi (2012)
• [40] Vissers (2014)
• [49] Chopade (2013)
• [52] Ismail (2012)
• [53] Chen (2011)
• [19] Karnwal (2012)
• [46] Huang (2013)
• [47] Fujinoki (2013)
• [48] Tripathi (2013)
• [54] Priyanka (2013)</td></tr>
</table>
</td></tr>
</table>

<table>
<tr>
<td><b>User Authentication</b></td>
<td>What form of user authentication is employed by the system? Many systems incorporate authentication protocols to identify legitimate users. Typically, strong user authentication involves puzzles such as CAPTCHA. Other methods include marking packets, but these can suffer from spoofing attacks. Below is a classification of recent solutions by whether or not the system uses puzzles for user authentication.
<table>
<tr><th>Yes</th><th>No</th></tr>
<tr><td>[47] Fujinoki (2013)
[48] Tripathi (2013)
[54] Priyanka (2013)
[40] Vissers (2014)</td><td>[39] Joshi (2012)
[49] Chopade (2013)
[52] Ismail (2012)
[53] Chen (2011)
[19] Karnwal (2012)
[46] Huang (2013)
[17] Yu (2014)
[41] Shtern (2014)
[8] Wang (2014)
[45] Jia (2014)</td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Response</b></td>
<td>What is the response method of the proposed system? The range of response methods is quite varied and can consist of several layers but can be grouped into categories. Responses include trace back techniques to discover the source of techniques, which can then be used to update filters. Other responses include techniques used to filter and drop packets and user requests, and those that migrate users to new VMs when current ones become compromised.
<table>
<tr><th>Trace back</th><th>VM Migration</th><th>Filters/Block lists/Dropped Packets</th></tr>
<tr><td>[19] Karnw (2012)
[39] Joshi (2012)</td><td>[41] Shtern (2014)
[8] Wang (2014)
[45] Jia (2014)
[47] Fujinoki (2013)</td><td>[46] Huang (2013)
[49] Chopade (2013)
[40] Vissers (2014)
[53] Chen (2011)
[52] Ismail (2012)</td></tr>
</table>
</td>
</tr>
</table>

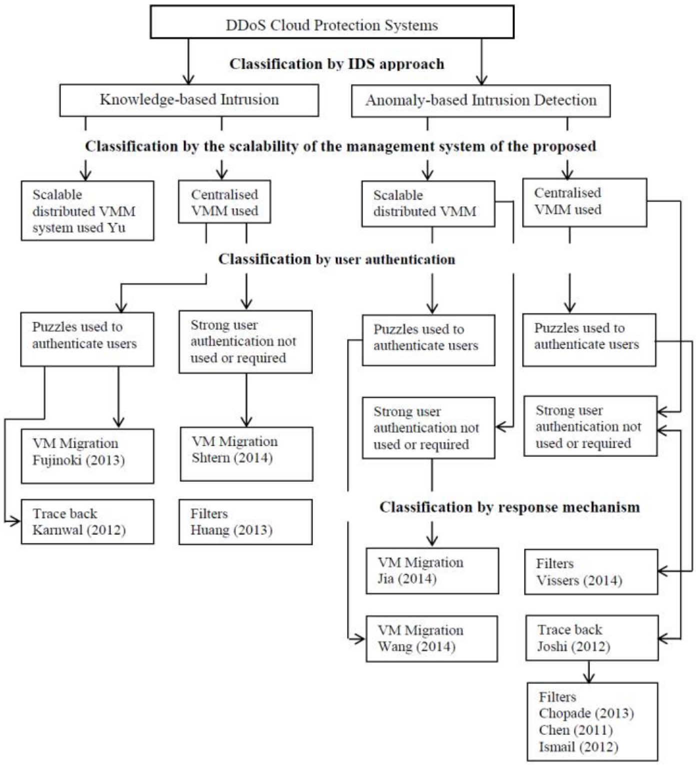

DDoS Cloud Protection Systems
Classification by IDS approach
Knowledge-based Intrusion
Anomaly-based Intrusion Detection
Classification by the scalability of the management system of the proposed
Scalable distributed VMM system used Yu
Centralised VMM used
Scalable distributed VMM
Centralised VMM used
Classification by user authentication
Puzzles used to authenticate users
Strong user authentication not used or required
Puzzles used to authenticate users
Puzzles used to authenticate users
Strong user authentication not used or required
Strong user authentication not used or required
VM Migration Fujinoki (2013)
VM Migration Shtern (2014)
Trace back Karnwal (2012)
Filters Huang (2013)
Classification by response mechanism
VM Migration Jia (2014)
Filters Vissers (2014)
VM Migration Wang (2014)
Trace back Joshi (2012)
Filters Chopade (2013) Chen (2011) Ismail (2012)

Fig. 2. DDoS cloud protection systems

IX. Conclusion

This paper examined the latest security issues in virtualisation technologies and IDPS to defend against DDoS attacks targeting cloud systems. Based on this comprehensive review, it is apparent that approaches to security across all reviewed areas share many common themes. These themes include:

1) Who and how will the security system be managed and monitored?

2) How will alerts be triggered?

3) How is the false alarm rate reduced?

4) What is the impact on operational overheads?

With virtualisation being a key underlying technology of the cloud computing paradigm it is understandable that there are a number of similarities between proposed systems. A number of these highlight the use of VMs as system management units and a few of these allow these to be generated in a similar way to user VMs in the cloud. This allows these systems to make use of the elasticity and scalability of the cloud paradigm to provide a more effective response to an attack and helps to reduce bottlenecks in the system.

A common response to an attack is to migrate users to new VMs through either physical or logical migration. This uses the strengths of VMs and helps to enforce the principles of isolation. Systems must be in place to remove compromised VMs to ensure that they are not migrated along with other users.

Commonly, the response of the system is designed based on non-attack or low-level attack conditions. This allows systems, even under test conditions to deliver the necessary messages to update filter protocols and perform other defensive manoeuvres. It must be considered, that under high stresses these systems may not operate in the same manner. Against a DDoS flooding attack, it may not be possible to update upriver filters effectively enough to reduce the intensity of the attack. Proposed systems must therefore test themselves under such strains so that their behaviours at these attack intensities can be observed.

The majority of defence systems focus on a particular type or point of attack against which they can be shown to be effective. The next step is to integrate these approaches to provide a more universal protection. When implementing this integration it is important that new vulnerabilities are not introduced into the system.

In the authors opinion, there are two research main research avenues to be followed. In the first, the intrusion is attempting to compromise VMs in order to launch a DDoS attack against a target that is external to the cloud. Once the intrusion is detected, a counter-measure is to be deployed, which in this case will be a calibrated firewall. Although this is may appear to be a simplistic fix to exiting protocols, it is not a solution that is widely adopted by current CSPs because it adds to their overheads, while not directly protecting their own infrastructure. The second considers a more traditional cloud intrusion where the target of the attack is the cloud or an element within the cloud itself. Resources relevant to this scenario will be based in a Eucalyptus cloud system.

REFERENCES

[1] M. Mackay, et al., "Security-oriented cloud computing platform for critical infrastructures," Computer Law & Security Review, vol. 28, pp. 679-686, 2012.

[2] C. Rong, et al., "Beyond lightning: A survey on security challenges in cloud computing," Comput. Electr. Eng., vol. 39, pp. 47-54, 2013.

[3] J. Li, et al., "CyberGuarder: A virtualization security assurance architecture for green cloud computing," Future Generation Computer Systems, vol. 28, pp. 379-390, 2012.

[4] J. Jang-Jaccard and S. Nepal, "A survey of emerging threats in cybersecurity," Journal of Computer and System Sciences, vol. 80, pp. 973-993, 2014.

[5] Galante J., et al. (2011, May 14, 2015). Sony network Breach shows Amazon Cloud's appeal for hackers. Available: http://bloomberg.com/news/2011-05-15/sony-attack-shows-amazon-s-cloud-service-lures-hackers-at-pennies-an-hour.html

[6] W. Wei. (2014, May 14, 2015). Sony Playstation Network Taken Down By DDoS Attack. The Hackers News. Available: http://thehackernews.com/2014/08/sony-playstation-network-taken-down-by_24.html

[7] M. Kumar. (2011, May 14, 2015). Cloud Computing Used to Hack Wireless Password. The Hackers News. Available: http://thehackernews.com/2011/01/cloud-computing-used-to-hack-wireless.html

[8] H. Wang, et al., "A moving target DDoS defense mechanism," Computer Communications, vol. 46, pp. 10-21, 2014.

[9] M. Darwish, et al., "Cloud-based DDoS attacks and defenses," in Information Society (i-Society), 2013 International Conference on, 2013, pp. 67-71.

[10] P. paganini. (2013, May 14, 2015). Cybercriminals using hijacked Cloud hosting accounts for targeted attacks. The Hackers News. Available: http://thehackernews.com/2013/06/cybercriminals-using-hijacked-cloud.html

[11] S. Khandelwal, "SNMP Reflection DDoS Attacks on the Rise. The Hackers News," 2014.

[12] A. Patel, et al., "An intrusion detection and prevention system in cloud computing: A systematic review," Journal of Network and Computer Applications, vol. 36, pp. 25-41, 2013.

[13] A. Rehman, et al., "Virtual machine security challenges: case studies," International Journal of Machine Learning and Cybernetics, vol. 5, pp. 729-742, 2014/10/01 2014.

[14] J. E. Smith and R. Nair, "The architecture of virtual machines," Computer, vol. 38, pp. 32-38, 2005.

[15] S. Subashini and V. Kavitha, "Review: A survey on security issues in service delivery models of cloud computing," J. Netw. Comput. Appl., vol. 34, pp. 1-11, 2011.

[16] B. Loganayagi and S. Sujatha, "Enhanced Cloud Security by Combining Virtualization and Policy Monitoring Techniques," Procedia Engineering, vol. 30, pp. 654-661, 2012.

[17] S. Yu, et al., "A Security-Awareness Virtual Machine Management Scheme Based on Chinese Wall Policy in Cloud Computing," The Scientific World Journal, vol. 2014, p. 12, 2014.

[18] Briscoe. (2011 Marinos: Digital Ecosystems in the Clouds: Towards Community Cloud Computing. Available: http://blog.ascens-ist.eu/wp-content/uploads/2011/03/xaas.png

[19] T. Karnwal, et al., "A comber approach to protect cloud computing against XML DDoS and HTTP DDoS attack," in Electrical, Electronics and Computer Science (SCEECS), 2012 IEEE Students' Conference on, 2012, pp. 1-5.

[20] R. Koch, et al., "Behavior-based intrusion detection in encrypted environments," Communications Magazine, IEEE, vol. 52, pp. 124-131, 2014.

[21] J. McHugh, "Intrusion and intrusion detection," Digital Object Identifier (DOI) 10.1007/s102070100001, pp. 14-35, : 27 July 2001 2001.

[22] S. M. Alqahtani, et al., "An Intelligent Intrusion Prevention System for Cloud Computing (SIPSCC)," in Computational Science and Computational Intelligence (CSCI), 2014 International Conference on, 2014, pp. 152-158.

[23] C. Modi, et al., "A survey of intrusion detection techniques in Cloud," Journal of Network and Computer Applications, vol. 36, pp. 42-57, 2013.

[24] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: A review," Applied Soft Computing, vol. 10, pp. 1-35, 2010.

[25] D. Moore, et al., "Inferring Internet denial-of-service activity," ACM Trans. Comput. Syst., vol. 24, pp. 115-139, 2006.

[26] R. Goel, et al., "Cloud Computing Vulnerability: DDoS as Its Main Security Threat, and Analysis of IDS as a Solution Model," in Information Technology: New Generations (ITNG), 2014 11th International Conference on, 2014, pp. 307-312.

[27] H. Beitollahi and G. Deconinck, "Analyzing well-known countermeasures against distributed denial of service attacks," Computer Communications, vol. 35, pp. 1312-1332, 2012.

[28] A. Volokyta, et al., "Secure virtualization in cloud computing," in Modern Problems of Radio Engineering Telecommunications and Computer Science (TCSET), 2012 International Conference on, 2012, pp. 395-395.

[29] L. Qian, et al., "An In-VM Measuring Framework for Increasing Virtual Machine Security in Clouds," Security & Privacy, IEEE, vol. 8, pp. 56-62, 2010.

[30] D. Williams, et al., "Security through Diversity: Leveraging Virtual Machine Technology," Security & Privacy, IEEE, vol. 7, pp. 26-33, 2009.

[31] O. Al-Jarrah and A. Arafat, "Network Intrusion Detection System using attack behavior classification," in Information and Communication Systems (ICICS), 2014 5th International Conference on, 2014, pp. 1-6.

[32] L. Laboratory. (2014, DARPA INTRUSION DETECTION EVALUATION. Available: http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/index.html

[33] C. Chun-Jen, et al., "NICE: Network Intrusion Detection and Countermeasure Selection in Virtual Network Systems," Dependable and Secure Computing, IEEE Transactions on, vol. 10, pp. 198-211, 2013.

[34] H. Jin, et al., "A VMM-based intrusion prevention system in cloud computing environment," The Journal of Supercomputing, vol. 66, pp. 1133-1151, 2013/12/01 2013.

[35] A. Alhomoud, et al., "Performance Evaluation Study of Intrusion Detection Systems," Procedia Computer Science, vol. 5, pp. 173-180, 2011.

[36] H. Mohamed, et al., "A collaborative intrusion detection and Prevention System in Cloud Computing," in AFRICON, 2013, 2013, pp. 1-5.

[37] A. Alazab, et al., "Using response action with intelligent intrusion detection and prevention system against web application malware," Information Management & Computer Security, vol. 22, pp. 431-449, 2014.

[38] Y. Lanjuan, et al., "Defense of DDoS attack for cloud computing," in Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on, 2012, pp. 626-629.

[39] B. Joshi, et al., "Securing cloud computing environment against DDoS attacks," in Computer Communication and Informatics (ICCCI), 2012 International Conference on, 2012, pp. 1-5.

[40] T. Vissers, et al., "DDoS defense system for web services in a cloud environment," Future Generation Computer Systems, vol. 37, pp. 37-45, 2014.

[41] M. Shtern, et al., "Towards Mitigation of Low and Slow Application DDoS Attacks," presented at the Proceedings of the 2014 IEEE International Conference on Cloud Engineering, 2014.

[42] P. C. Senthilmahesh, et al., "DDoS Attacks Defense System Using Information Metrics," in Proceedings of the Third International Conference on Trends in Information, Telecommunication and Computing. vol. 150, V. V. Das, Ed., ed: Springer New York, 2013, pp. 25-30.

[43] R. Mathew and V. Katkar, "Survey of low rate DoS attack detection mechanisms," presented at the Proceedings of the International Conference &#38; Workshop on Emerging Trends in Technology, Mumbai, Maharashtra, India, 2011.

[44] Y. Tang, "Countermeasures on Application Level Low-Rate Denial-of-Service Attack," in Information and Communications Security. vol. 7618, T. Chim and T. Yuen, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 70-80.

[45] J. Quan, et al., "Catch Me If You Can: A Cloud-Enabled DDoS Defense," in Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on, 2014, pp. 264-275.

[46] V. S. Huang, et al., "A DDoS Mitigation System with Multi-stage Detection and Text-Based Turing Testing in Cloud Computing," in Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on, 2013, pp. 655-662.

[47] H. Fujinoki, "Dynamic Binary User-Splits to Protect Cloud Servers from DDoS Attacks," presented at the Proceedings of the Second International Conference on Innovative Computing and Cloud Computing, Wuhan, China, 2013.

[48] S. Tripathi, et al., "Hadoop Based Defense Solution to Handle Distributed Denial of Service (DDoS) Attacks," Journal of Information Security, 2013.

[49] S. S. Chapade, et al., "Securing Cloud Servers Against Flooding Based DDOS Attacks," in Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013, pp. 524-528.

[50] R. Latif, et al., "Distributed Denial of Service (DDoS) Attack in Cloud-Assisted Wireless Body Area Networks: A Systematic Literature Review," J. Med. Syst., vol. 38, pp. 1-10, 2014.

[51] L. Chi-Chun, et al., "A Cooperative Intrusion Detection System Framework for Cloud Computing Networks," in Parallel Processing Workshops (ICPPW), 2010 39th International Conference on, 2010, pp. 280-284.

[52] M. N. Ismail, et al., "Detecting flooding based DoS attack in cloud computing environment using covariance matrix approach," presented at the Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication, Kota Kinabalu, Malaysia, 2013.

[53] C. Qi, et al., "CBF: A Packet Filtering Method for DDoS Attack Defense in Cloud Environment," in Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on, 2011, pp. 427-434.

[54] N. Priyanka, et al., "Enhanced CBF Method to Detect DDoS Attack in Cloud Computing Environment," International Journal of Computer Science Issues, IJCSI, vol. 10, pp. 142-146, 2013.

Optimal Design of PMSA for SBW Application

Rached BEN MEHREZ
Research Unit on Signals and Mechatronic Systems SMS,
Ecole National d’Ingénieurs de Carthage,
45 Rue des Entrepreneurs 2035 -Charguia2 -Tunis -Tunisie

Lilia EL AMRAOUI
Research Unit on Signals and Mechatronic Systems SMS,
Ecole National d’Ingénieurs de Carthage,
45 Rue des Entrepreneurs 2035 -Charguia2 -Tunis -Tunisie

***Abstract*—In this paper a new topology of Permanent Magnet Synchronous Actuator (PMSA) is used for steer-by-wire application. The magnetic field patterns are determined from finite element modeling, for different rotor positions and supply currents, using FEMM software. The designed actuator geometric is, then, optimized using Genetic Algorithm in order to ameliorate its electromagnetic characteristics, and its resulting torque. Finally, a thermal analysis is achieved for the initial and the optimized actuators. The obtained results show a clear improvement of the actuator electromagnetic characteristics and heat distribution.**

Keywords—Genetic Algorithms (GA); Permanent Magnet Synchronous Actuator (PMSA); finite elements analysis; Steer-By-Wire application (SBW); Thermal study; Optimization

I. Introduction

Permanent Magnet Synchronous Actuators (PMSA) have become more attractive because they respond well to new technology requirements [1]. The renewed interest for these machines is due, in large part, to their excellent dynamic characteristics, low loss and their high specific torque, making them better suited to industrial applications requiring electrical drive control position or speed [2].

PMSA are suitable for low speed and high torque applications. Our contribution consists on the design of an actuating system for Steer-By-Wire (SBW) application; based on PMSA. The SBW drive system does not include gearbox, the electric machine being directly driven by the motion control system. The removal of the gearbox that typically represents a source of important losses and involves high maintenance costs is very attractive and permits the simplification of the drive cinematic conversion chain [3].

The electric machine in case of high torque applications should operate at low rotational speeds and thus it should have a large number of magnetic poles.

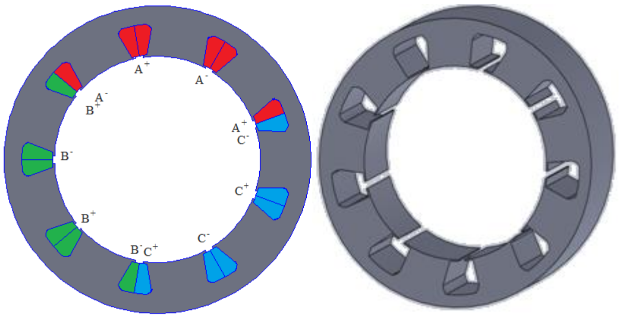
In the first section, the finite element model of PMSA is analyzed. Then, the finite element model of the actuator designed is optimized using a genetic algorithm. Finally electromagnetic performances of the actuator are analyzed and interpreted.

II. Study Actuator

A Permanent magnet synchronous actuator (PMSA) has been applied to the performance improvement of electric power steering. Because PMSA have many advantages such as high efficiency and high torque per rotor volume, they are especially suitable for automotive applications, where space and energy savings are critical [4].

A. *Structure presentation*

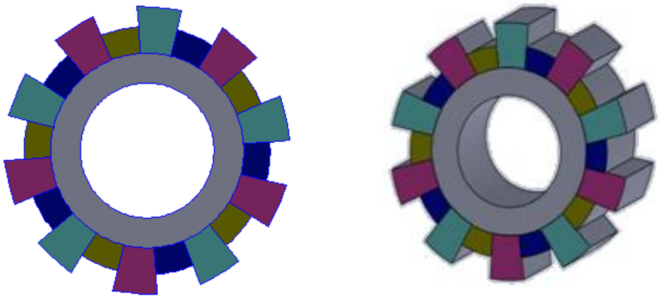
The designed actuator shown in Fig.1 and Fig.2 has nine poles on the stator and ten poles on the rotor, with permanent magnets disposed to the rotor. Thus, for the same flux density in the air gap of the ferrite magnets with low costs can be used [5]. There is a rotary Permanent Magnet Synchronous Motor, whose polarization is substantially Albach type [6] containing the supply coils; Fig.1.a shows the Coils distribution.



a. Two-dimensional view b. Perspective view

Fig. 1. Stator structure

The rotor presented by Fig.2 is composed of a ferromagnetic ring supporting 10 magnets which has radials magnetized magnet and 10 longitudinally magnetized ones, alternated.



a. Two-dimensional view b. Perspective view

Fig. 2. Rotor structure

The designed actuator dimensions are given in Table VI.

B. *Magnetic characteristics*

The magnetic circuit of the rotor and the stator are made from iron-silicium Fe-Si, the stack of the stator is made of steel. Each winding consists on a coil of 63 turns, the magnetization characteristic of the Fe-Si material is shown in Fig.3.

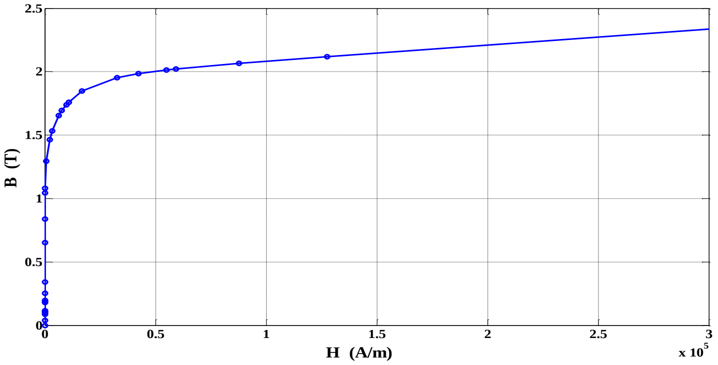


Fig. 3. Fe-Si magnetization characteristic

The ramenante induction of the magnet is equal to 1 tesla.

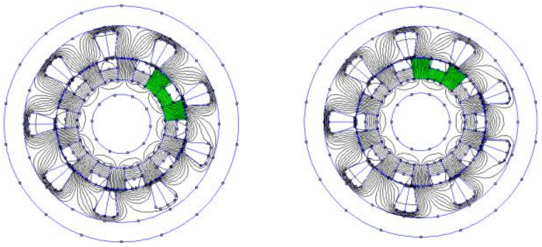
III. Finite Element Modeling of the Studied Actuator

After material characteristic assignment, meshing and solving the magnetic problem under FEMM, flux linkage, back emf, current wave form and instantaneous torque are computed.

A. Flux linkage

The induction line in the actuator is shown in Fig.4. These lines are generated by the permanent magnets for effective current density equal to 0 A/mm^2.

Figure 4 shows the distribution of induction lines in the actuator, for two considered positions; the first position corresponds to junction position of phases A, the second one is shifted by 5° from the first one. The induction distribution depends on the actuator moving part position.



a. First position b. Second position

Fig. 4. Induction line distribution in the actuator

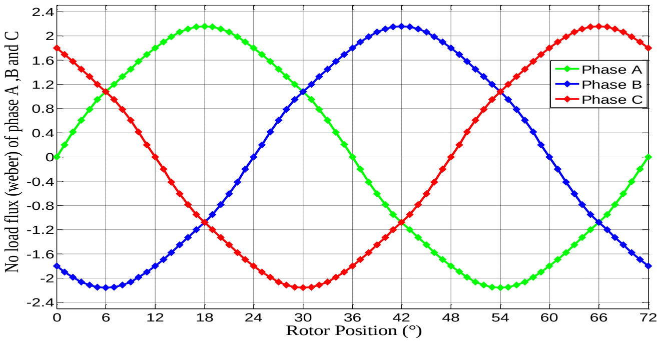


Fig. 5. No load flux of phases A, B and C according to rotor position

The no load flux developed by phases A, B and C are presented in Fig. 5. These characteristics are sinusoidal and shifted from each other by a third of the mechanical period. Thus, the corresponding flux can be approximated by the equation (1):

$$\begin{cases} \Phi_{a\ exi} = \sqrt{2}\,\Phi_{eff}\,\sin\left(\dfrac{N_r}{2}\theta\right) \\ \Phi_{b\ exi} = \sqrt{2}\,\Phi_{eff}\,\sin\left(\dfrac{N_r}{2}\theta - \dfrac{2\Pi}{3}\right) \\ \Phi_{c\ exi} = \sqrt{2}\,\Phi_{eff}\,\sin\left(\dfrac{N_r}{2} - \dfrac{4\Pi}{3}\right) \end{cases} \qquad (1)$$

B. Electromotive Force

Referring to the laws of Lenz, the no load back emf are given by (2):

$$\begin{cases} e_a = \dfrac{d\left(\Phi_{a\ exi}\right)}{dt} = \dfrac{d\left(\Phi_{a\ exi}\right)}{d\theta}\dfrac{d\theta}{dt} \\ e_b = \dfrac{d\left(\Phi_{b\ exi}\right)}{dt} = \dfrac{d\left(\Phi_{b\ exi}\right)}{d\theta}\dfrac{d\theta}{dt} \\ e_c = \dfrac{d\left(\Phi_{c\ exi}\right)}{dt} = \dfrac{d\left(\Phi_{c\ exi}\right)}{d\theta}\dfrac{d\theta}{dt} \end{cases} \qquad (2)$$

According to equations (1) and (2) the characteristics of back emf, depending on mechanical position, are presented in Fig.6, for a speed of Ω=10 rad/s

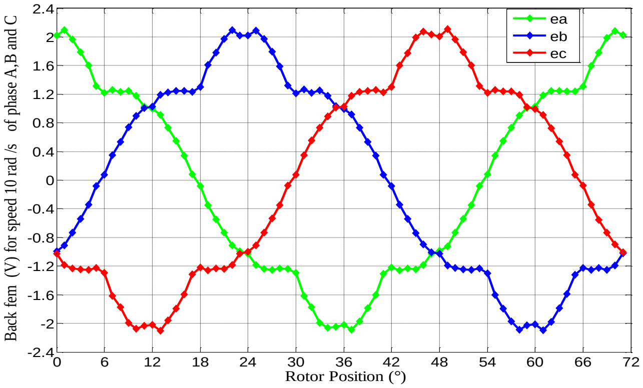


Fig. 6. Back emf for speed of 10 rad /**s**

These characteristic are triangular wave form and shifted from each other by a third of the mechanical period.

C. Current supply wave form

The supply current characteristics in this case are put in phase with the back fem, which is governed by the equations (3):

$$\begin{cases} i_a = \sqrt{2}\dfrac{JS_{slot}k_b}{N_{spir}}\cos(\dfrac{N_r}{2}\theta) \\ i_b = \sqrt{2}\dfrac{JS_{slot}k_b}{N_{spir}}\cos(\dfrac{N_r}{2}\theta - \dfrac{2\Pi}{3}) \\ i_c = \sqrt{2}\dfrac{JS_{slot}k_b}{N_{spir}}\cos(\dfrac{N_r}{2}\theta - \dfrac{4\Pi}{3}) \end{cases} \qquad (3)$$

The current wave forms of phases A, B and C according to mechanical rotor position are shown in Fig. 7.

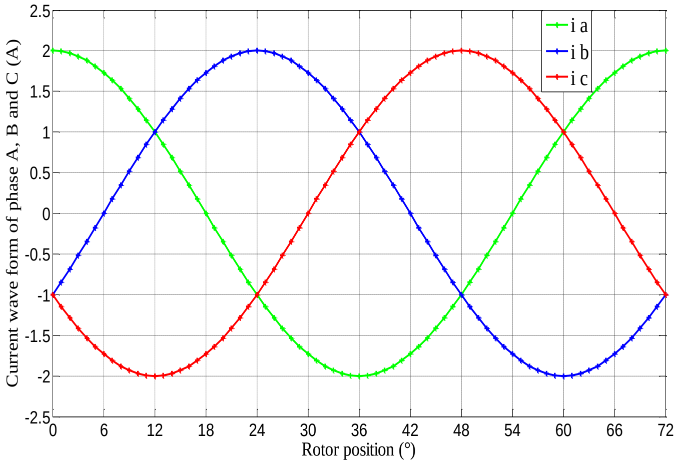


Fig. 7. Current wave form of phases A, B and C according to rotor position

D. Instantaneous torque

For different supply current densities varying from 0 to 62A/mm^2 the electromagnetic torque is computed over one mechanical period and for constant current densities, according the coenergy W' from the expression (4)

$$C = \left.\frac{\partial W'}{\partial \theta}\right|_{i=cst} \qquad (4)$$

For determining the torque at a given position, it is necessary to carryout simulations for two near mechanical positions, while the current is kept constant. The mechanical angular variation $\Delta\theta$ between the two simulation positions gives an approximate calculation of torque performing a derivative of coenergy from (5):

$$C = \left.\frac{W'(\theta+\Delta\theta) - W'(\theta)}{\Delta\theta}\right|_{i=cst} \qquad (5)$$

The obtained response surface of torque, according to rotor position and current densities, is shown on Fig. 8.

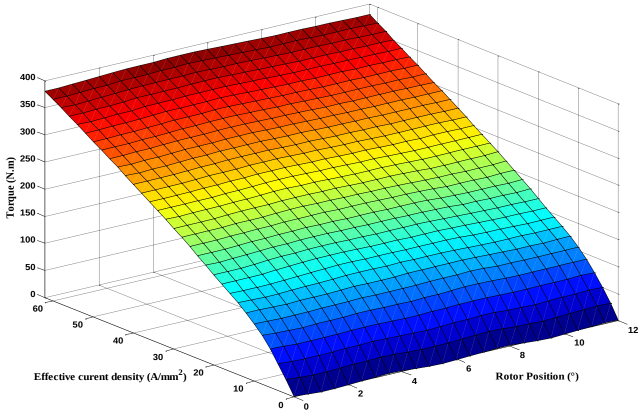


Fig. 8. Response surface of torque

The response surface of torque shows that the electromagnetic torque produced by PMSA, for a given current, is constant over the mechanical period.

IV. Optimization of the Designed Permanent Magnet Synchronous Actuator

The structure of the designed actuator is, in this section, optimized in order to improve the actuator electromagnetic performances, by use of the finite elements model as virtual prototype and genetic algorithm as optimization software.

The objective function is the maximization of electromagnetic torque.

A. Optimization problem

Genetic Algorithms (GA) are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics. As such, these represent an intelligent exploitation of a random search used to solve optimization problems [7].

The optimization problem is described by (P) from the expression (6):

$$P: \begin{cases} Objective\ \ function: Max(C) \\ parametres: \\ \qquad 25\text{mm} \le R_{\text{shaft}} \le 40\text{mm} \\ \quad 102.677\ \text{mm}^2 \le S_{\text{slot}} \le 551.931\ \text{mm}^2 \\ \\ under\ constraints \\ \qquad R_{\text{exts}} = 100mm \\ \qquad e = 0.2mm \\ \qquad L_a = 100mm \\ \qquad J = 40A/mm^2 \end{cases} \qquad (6)$$

The parameters of G.A. are given by table I.

TABLE I. G.A. PARAMETERS

Parameters	Value\ types
Generation number	29
Population number	10
Crossover probability	0.8
Mutation probability	0.2
Crossing type	Scattered
Selection type	Stochastic uniform
Mutation type	Uniform

B. *Results of the G.A. optimization*

The following table shows the parameters of the initial and the optimized structure of the studied actuator.

TABLE II. PARAMETERS OF INITIAL AND OPTIMIZED STRUCTURES

Parameter	Initial structure	Optimized structure
S_{slot}	466.688 mm^2	332.054 mm^2
R_{shaft}	30 mm	31.15 mm
C	273.57 N.m	327.15 N.m

Figure 9 shows the resulting characteristics of torque developed by the initial and the optimized structures over one mechanical period.

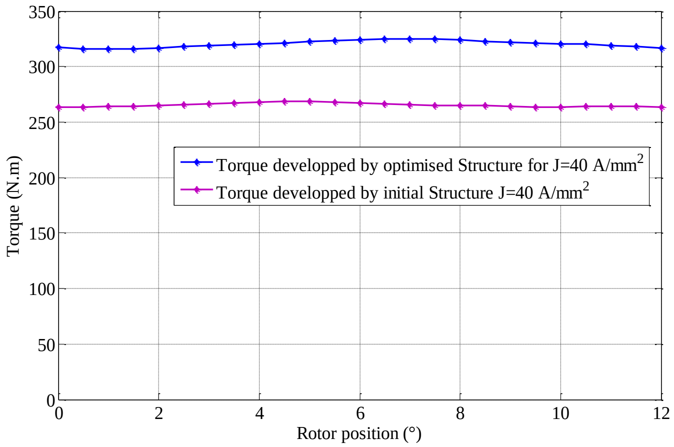


Fig. 9. Torque developed by initial and optimized PMSA structures

Referring to (3) and for different supply current densities varying from 0 to 62A/mm2 the electromagnetic torque of the optimized PMSA structure is computed over one mechanical period, Fig. 10.

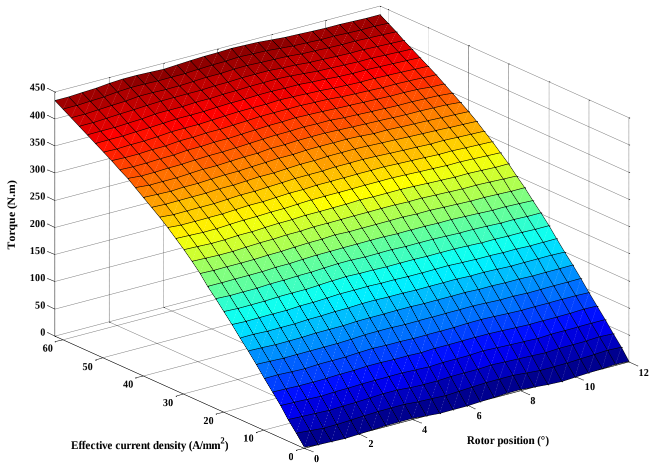
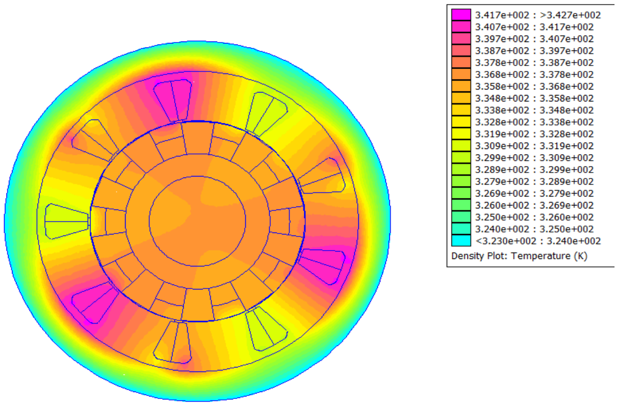


Fig. 10. Torque response surface of the optimized Actuator

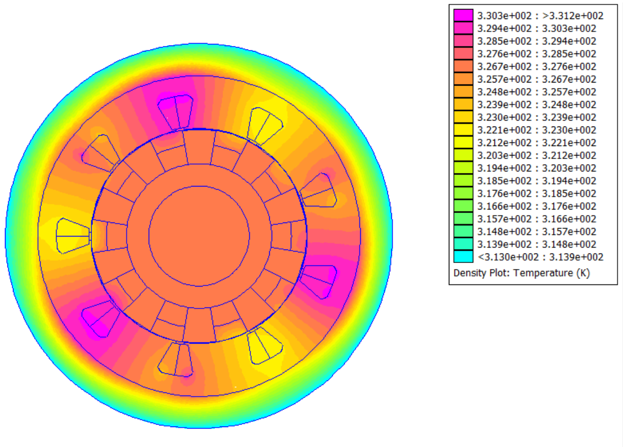
V. THERMAL STUDY OF INITIAL AND OPTIMIZED ACTUATORS

The temperature rise is caused by the operation of rotating machines is an unavoidable natural phenomenon. Computer Aided Design (CAD) systems are used to describe geometrical models, to manipulate data and to display the results. Therefore, CAD systems with thermal analysis and computational modules become very powerful tools to investigate the thermal distribution within the electronics package, [8, 9].

An analysis of the temperature distribution, Fig. 11, is required to ensure that there is no concentration of heat that could cause damage during operation or demagnetization of permanent magnet.



(a)



(b)

Fig. 11. Temperature distribution into actuator: (a) Initial structure, (b) Optimized structure

The Thermal properties of materials are given in table III.

TABLE III. THERMAL PROPERTIES OF THE ACTUOR MATERIALS

Segment	Material	α	β
Permanent magnet	NdFeb	3.11	9
Stator and shaft	Silicon Steel	3.77	20e^{-30}
Stator Slot	Copper	3.40	360
Air gap	Air	0.0012	0.025

With: α: Specific heat capacity (MJ/m^{3}K) and β: Thermal Conductivity (W/mK)

Figure 12 shows the temperature cartography into initial and optimized structure. The temperature reached into the initial structure (from 323 to 341° K) is higher than that obtained for the optimized structure (from 313 to 330° K).

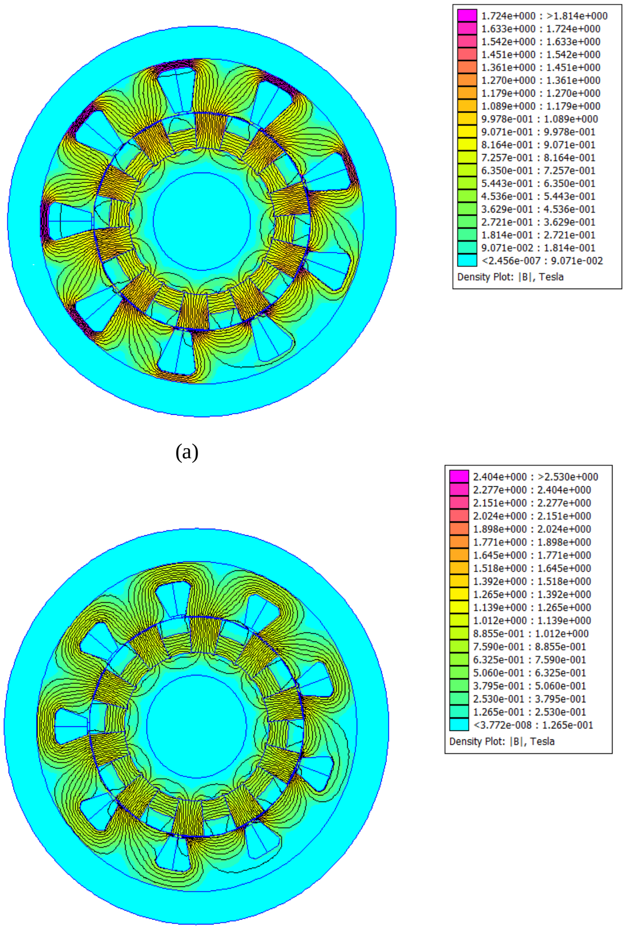
VI. RESULTS' COMPARISON

The following table summarizes the finite element results obtained from the initial and the optimized structures.

TABLE IV. PARAMETERS AND ELECTROMAGNETIC CHARACTERISTICS OF THE INITIAL AND OPTIMIZED STRUCTURES

	Initial structure	Optimised structure
Nombre of phase	3	3
Outer radius of the rotor (mm)	68.3	68.3
Inner radius of the rotor (mm)	**30**	**31,15**
Outer radius of the stator(R_{exts})	100	100
Inner radius of the stator (mm)	66.3	66.3
Winding coefficient (K_b)	0.8	0.8
Phase turn number (N_{spir})	**63*3**	**44*3**
Air gap (e) (mm)	0.2	0.2
Number of stator teeth (Ns)	9	9
Number of rotor teeth (Nr)	10	10
Rotor iron volume (cm^3)	**353.370**	**331.156**
Active length (L_a)	100	100
Cross section of slot (mm^2)	**466.688**	**332.054**
Maximum torque value produced (N. m)	**377**	**436**
Torque gains	**15.64%**	
Heat gains	**3 %**	
Rotor iron gains	**6.71 %**	

Figure 12 shows the cartographies of the magnetic induction field corresponding to the initial and optimized actuator.



(a)

(b)

Fig. 12. Magnetic induction field: (a) Initial actuator, (b) Optimized actuator

According to the results shown on Fig. 12, it can be seen that both stator and rotor poles are much more saturated in the initial structure, than, in the optimal one.

The results' comparisons, table IV, show a torque gain of 15.64 % in the optimized structure, as well as in the phase turn number, which decreases from 63 to 44. Furthermore, the rotor iron volume gain is of 6.71 % and the heat level gain is of 3%.

VII. CONCLUSION

A new structure of permanent magnet actuator is designed for a steer-by-wire application. It is modeled and characterized under the finite element software FEMM environment. In the first part, the actuator initial structure is analyzed. In the second part, this structure is optimized using genetic algorithm. Finally a thermal analysis of initial and optimized structures is achieved.

The obtained results show clear improvement of torque, phase turn number, rotor iron volume and heat level in the optimized structure according to the initial one.

In future works the overall dynamic behavior of SBW direction actuated by the studied PMSA will be optimized.

Appendix

C	Torque (N.m).
J	Effective current density (A/mm^2).
S_{slot}	Slot section (mm^2) .
N_{spir}	Number of turns per phase.
N_s	Number of stator tooth.
N_r	Number of rotor tooth.
K_b	Coefficient of winding.
θ	Mechanical position (deg).
$i_{a,b,c}$	Phase current a, b, c (A).
ϕ_{mi}	No load flux of phase i.
ϕ_{ij}	Linkage flux of phases i and j .
e_i	Back force of phase *i*.
ϕ_{eff}	Rms value of flux .
$\phi_{i\ exi}$	Excitation flux of phase *i*.
R_{shaft}	Radius of shaft .
e	Air-gap.
L_a	Active length.
$W^{'}$	Coenergy.

References

[1] Ait-Oufroukh, N. Messaoudene, K. .Mammar S., "Dynamic model of steer-by-wire system for driver handwheel feedback", 10th IEEE International Conference on Networking, Sensing and Control (ICNSC),vol., no., pp.780,785, 10-12 April 2013.

[2] B. Singh B.P singh, S. Dwived, " A state of Art on Different Configuration of Permanent Magnet Brushless Machines "IE(I) Journal-EL,pp.63-73,Vol87,June 2006.

[3] T. Tudorache, M. Popescu, "Optimal Design Solutions for Permanent Magnet Synchronous Machines," Advances in Electrical and Computer Engineering, vol. 11, no. 4, pp. 77-82, 2011.

[4] Hu Zhang, Jianwei Zhang, Konghui Guo, "An adaptive predictive current controller for Electric power steering system with Permanent Magnet Synchronous Motor," Transportation Electrification Asia-Pacific (ITEC Asia-Pacific), IEEE Conference and Expo , vol., no,pp.1,6,Sept.2014.

[5] Cao, W., B. C. Mecrow, G. J. Atkinson, J. W. Bennett, and D. J. Atkinson, "Overview of electric motor technologies used for more electric aircraft (MEA)," IEEE Transactions on Industrial Electronics, Vol. 59, No. 9, p. 3523-3531, 2012.

[6] S.M. Jang, S.S. Jeong, D.W. Ryu, S.K. Choi, " Design and analysis of a high speed slotless PM machine with Halbach array ", IEEE Transactions on Magnetics, vol. 37, no. 4, pp. 2827-2830, July 2001

[7] S. Singh, S Agrawal and D.V. Avasthi, "Optimization of design parameters of glazed hybrid photovoltaic thermal module using genetic algorithm," Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH) vol. no.20, pp.405,410, 28-29 Nov. 2014.

[8] A. Shah, B. G. Sammakia, H. Srihari, and K. Ramakrishna, "A numerical study of the thermal performance of an impingement heat sink-fin shape optimization," IEEE Trans. Components and Packaging Technologies, vol. 27, Issue 4, pp. 710 – 717, Dec. 2004.

[9] J. R. Culham and Y. S. Muzychka, "Optimization of plate fin heat sinks using entropy generation minimization", IEEE Trans. Components and Packaging Technologies, vol. 24, Issue 2, pp. 159 – 165, June 2001.

Experimental Study of the Cloud Architecture Selection for Effective Big Data Processing

Evgeny Nikulchev
Moscow Technological Institute,
National Research University – Higher School of Economics
Moscow, Russia

Dmitry Biryukov
Moscow State Technical University of Radio Engineering, Electronics and Automatics
Moscow, Russia

Evgeniy Pluzhnik
Moscow Technological Institute
Moscow, Russia

Oleg Lukyanchikov
Moscow Technological Institute,
Moscow State Technical University of Radio Engineering, Electronics and Automatics
Moscow, Russia

Simon Payain
Moscow Technological Institute
Moscow, Russia

***Abstract*—Big data dictate their requirements to the hardware and software. Simple migration to the cloud data processing, while solving the problem of increasing computational capabilities, however creates some issues: the need to ensure the safety, the need to control the quality during data transmission, the need to optimize requests. Computational cloud does not simply provide scalable resources but also requires network infrastructure, unknown routes and the number of user requests. In addition, during functioning situation can occur, in which you need to change the architecture of the application — part of the data needs to be placed in a private cloud, part in a public cloud, part stays on the client.**

Keywords—Cloud Infrastructure; Big Data; Distributed Databases; Hybrid Clouds

I. Introduction

Modern applications operate on large volumes of data that may reside in different stores. Cloud computing and cloud data storage are rapidly evolving, which gives advantages in performance due to parallel computing, the use of virtualization technology, scaling of computing resources and providing access to data via a web interface. Therefore, the actual task is to migrate existing systems and databases (DB) to the cloud.

Currently many developers and users are concerned about full advantage of cloud services. However, it is hard to tell in advance if a certain feature would be effective. Quite often new application features change the data structure. Furthermore, sometimes even a small modification can attract a large number of users, requiring database structure optimization to handle these large amounts of data. Currently the Big Data causes many problems for developers, since classic theories of query optimization only consider structure distribution and do not take inquiries depth into account. A new optimization parameter is introduced, called the data capacity with corresponding access and data transfer time. In this case normalized structures don’t have to be optimal, based on the transfer time. This requires novel development techniques for data analysis, given the constantly growing amount of data and changing data structures. However, so far migration of existing systems to the cloud only creates problems. Security issues of access to data and guarantee of service can be solved by using a hybrid cloud - part of the data (processing of queries that require large computational resources and not confidential data) is placed in the public cloud services, and the remaining data — in the private cloud or local network infrastructure [1]. However, in this case specialized design principles of cloud systems are yet to be developed. In theoretical terms this problem was considered in [2–4]. There are few solutions for specific applications [5–7].

Complexity of construction techniques for distributed databases in a hybrid cloud is that it is impossible to estimate the parameters of algorithms and query performance: in each case the acquired amount of cloud resources, such as virtual machines, is different; routes and characteristics of communication channels are unknown. Optimization of the volume and types of resources are another important task. Therefore, at present, in the absence of developed general principles and methods, experiments represent the only way to study the effectiveness of design decisions for this research field.

There was created an experimental stand (ES) that simulates the work with hybrid storage. Stand itself and some experimental results obtained with it are described in [8, 9], see figure 1. Employed software VMWare vCloud allows you to organize at all levels. VMware ESXi is used on two servers to create a cloud in the ES. Management system VCenter and application VMware vCloud Director are deployed. In ES there are more than 15 physical Cisco 29 switches and routers Series 26 and Series 28, as well as virtual switches Nexus. System based on ES allows to simulate routes of access to the data, to converge and diverge channels (can be done dynamically).

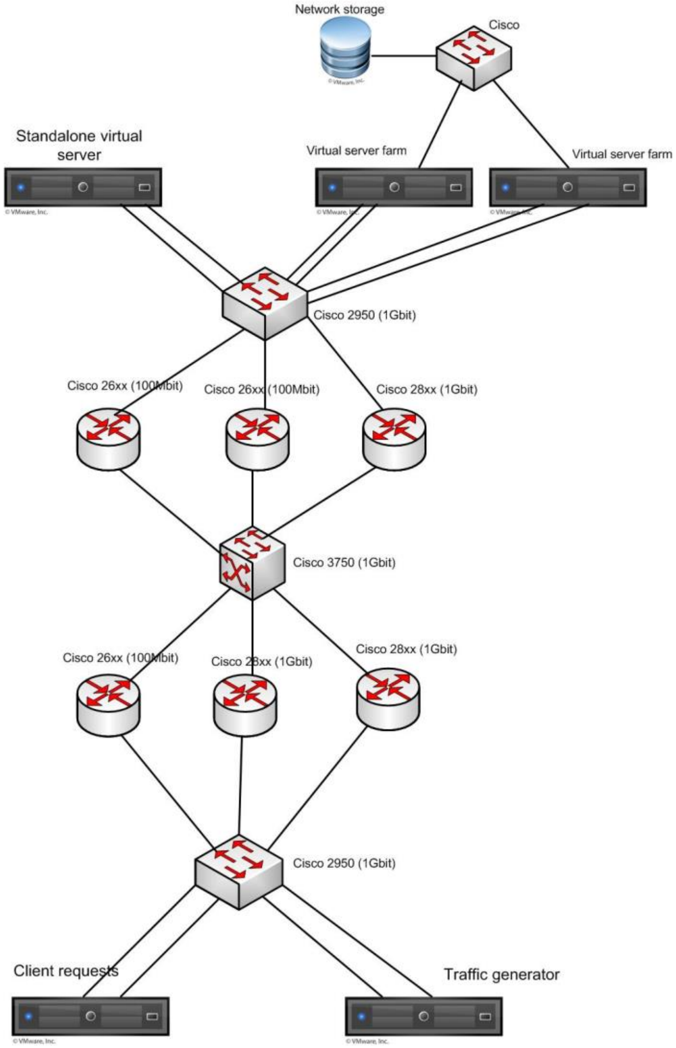


Fig. 1. Experimental installation

II. Experiments

For designed experimental setup, that emulates a hybrid cloud, experiment was conducted. Two types of database partitioning between public and private parts were examined. This is an important task for big data processing — in the process of operation and development of applications using the database with large volumes, including structured information, it is often required to transfer part of the data in the public cloud with a large number of resources, while not violating security requirements (i.e., leaving a portion of the data in the private cloud).

The aim of this experiment is to determine the efficiency of separation of the DB into 2 parts: public and private. For the experiment 3 compute nodes were prepared, the overall structure of which is shown in figure 2:

- Public DBMS server, which is more powerful (dual-core processor, 4 GB of RAM).
- Private DBMS server, which is less powerful (single-core processor, 2 GB of RAM).
- A client that makes requests to the published server using specialized software.

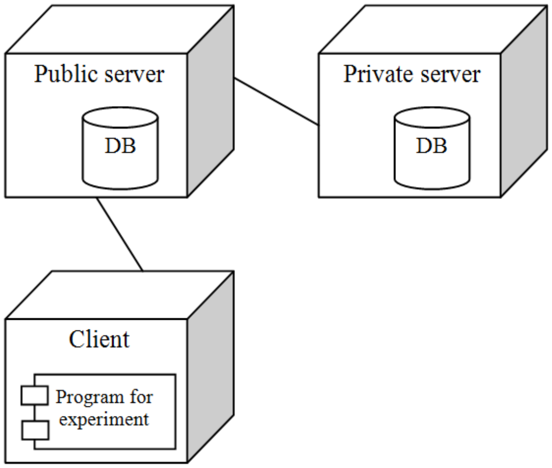


Fig. 2. The structure of the experiment

For the experiment there was used part of the database included in the educational process at the University (figure 3).

Table "Students" contains the following information about students:

- "id_stud" - unique number of the student, it is the primary key;
- "fio" - surname, name and second name of the student;
- "birthday" - the date of student’s birth;
- "agv_score" - the average score;
- "group" - the group number, which includes the student; is an external key of the table "Groups".

Table "Groups" associated one-to-many with table "Students" contains the following information about groups:

- "id_group" - unique group number;
- "name" - the name of the group.

Table "Lections" associated many-to-many with table "Students" via table "visit", contains following information about lectures:

- "id_lection" - unique number of lectures;
- "subject" - the number of the held object;
- "room" - the cabinet number (the audience);
- "theme" - is the theme of the lecture;
- "date" - the date of the lecture.

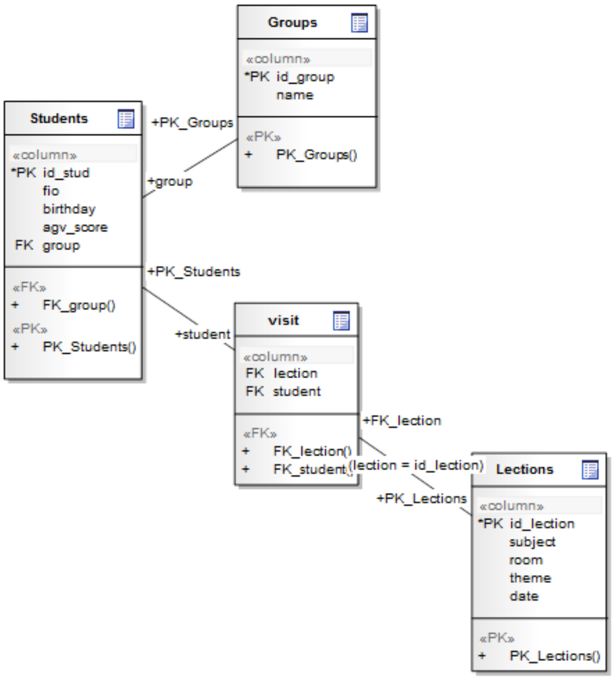


Fig. 3. The database scheme

Table "visit" includes the attendance of students.

- "id_student" - the number of student. Is external to the key of the table "Students";
- "id_lection" – the number of the lecture. Is external key to the table "Lections".

For this experiment the database was filled with random data:

Table "Students" - 3 000 000 records;

Table "Groups" - 100 records;

Table "Lections" - 100 000 records;

Table "visit" - 100 000 records.

The results of the query fetching all data from table "students" (adding data from tables associated one-to-many and many-to-many) were compared to test the effectiveness of the separation of the database into public and private parts.

In the first case, the entire database was on a public server, to retrieve the data following query was used:

*select * from students.students*

left join students.groups on students."group" = groups.id_group

*left join (select * from students.visit*

left join students.lections on visit.id_lection = lections.id_lection) t1

on students.id_stud = t1.id_student.

In the second case the part that relates many-to-many with table "Students" was placed on a private server (figure 4). Data was obtained by the function PostgreSQL dblink, which allows to perform the query to another DBMS. Request in the second case:

*select * from students.students*
left join students.groups on students."group" =
groups.id_group
*left join (select * from dblink('hostaddr=xxx.xxx.xxx.xxx*
port=xxxx dbname=… user=… password=…', 'select
id_student,lections.id_lection,id_subject,date,room,theme from
students.visit

left join students.lections on visit.id_lection =
lections.id_lection') as t(id_student INTEGER,id_lection
INTEGER,id_subject INTEGER,date DATE,room
INTEGER,theme TEXT)) as t1
on students.id_stud = t1.id_student

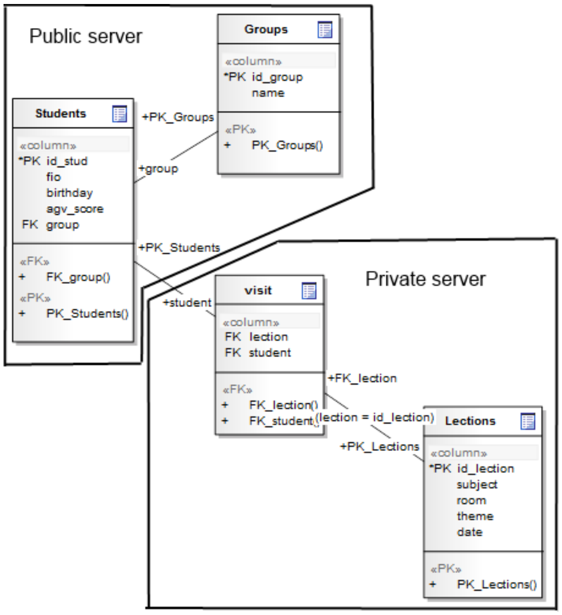


Fig. 4. Split the database between the public and private servers

Using such functions as dblink (which allows to perform queries to another DBMS), makes it possible to bring part of the database to another DBMS with changes to the query without need of modifying client application.

Effectiveness evaluation was performed using specialized software on the client, which emulates operation of the client application. To do this, through a small random amount of time the application has been generating queries, getting data about several students.

In the first case, when the entire database is on a public server, for 10 minutes 130 requests were performed. Time efficiency is presented in figure 5 (on the x-axis is the number of the query, on the y-axis is the query execution time in seconds).

In the second case, when the database is divided between public and private servers, for 10 minutes 126 requests were performed. Time efficiency is presented in figure 6 (on the x-axis is the number of the query, on the y-axis is the query execution time in seconds).

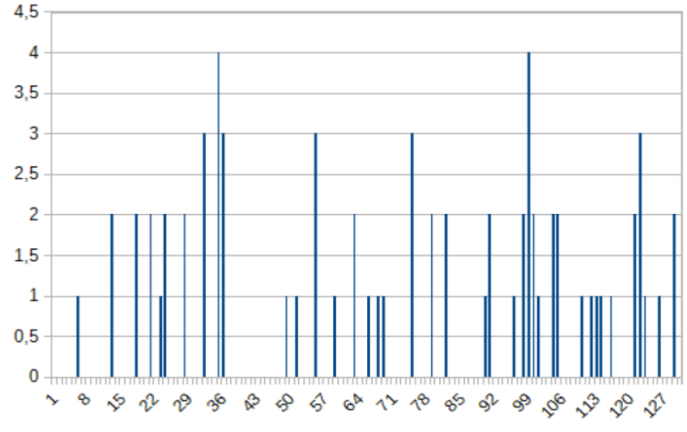


Fig. 5. Time efficiency when performing queries on the same public server

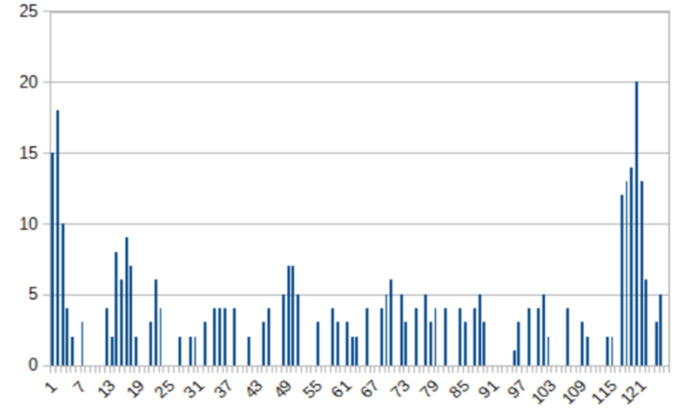


Fig. 1. Time efficiency when performing queries to public and private servers

III. Discussion

Main feature of the application is the intermediate layer that implements connection of user requests to the location of distributed data. Presence of unknown destination switching when using public cloud and mobile client makes it impossible to estimate the time of the algorithms. Here is why it's advisable to use software technology to control all stages of the system. However, hybrid infrastructure has many positive aspects of cloud computing: scalability, virtualization and also (due to the distribution of data) safety and security of data; designing information systems in the cloud has following problems:

- Impossible to assess the execution of individual queries and stream query;
- No general principles for designed systems with large amounts of data (Big data);
- Considerable amount of educational and scientific data is semistructured (XML);
- No technology for migration of database to the cloud – need to rewrite the code when moving to hybrid cloud;
- No commonly accepted principles for virtualization management allocation of resources in the cloud;
- The limitations associated with using an obsolete protocol (TCP).

The task was to account these features and develop the technology that creates applications in a hybrid cloud.

Within these limitations, the principles of development that provide guaranteed quality and functioning of the application were developed:

1) The system design should be based on a preliminary study on the simulation and experimental models.

2) It is necessary to control the main parameters of the infrastructure.

3) The use of object-oriented technology modifications of database design.

4) Technology should provide the flexibility of system structure, data volume, number of requests.

The result shows that in the second case, when the database is divided, the average query execution time is much higher than with solid database. It means that the separation of the database on the public and private parts adversely affects the performance of the system, and only has advantages from the viewpoint of safety.

Distributed data complicates the development of software, making it difficult and time-consuming to use common programming techniques. Despite development of technologies such as .Net and Qt, developers eventually have to operate SQL queries and clearly prescribe access to the distributed data. In the context of widespread object-oriented development methodology and application systems, with relational DBMS having dominant position in the market, advisable solution is to use intermediate software that provides necessary object-oriented interface to the data stored under control of a relational DBMS. To communicate with developed relational data objects there was used Object Relational Mapping (ORM) [10]. The essence of this technology is in accordance of programming entity to relational database object: each field of a table is assigned to a class attribute of an object.

The basic steps are the following:

1) Determination of basic structure of physically distributed in the hybrid cloud data.

2) Development of relational database structure.

3) Development of methods for data processing based on physical location of the data.

4) Creating classes of objects, including data and methods for their treatment.

5) Modification of the structure as a result of experimental study on the simulation bench.

6) Changing methods of processing inheritance

Depending on the task, system can be restructured to increase the speed of the most common queries, or to perform the most demanding requests in the public cloud.

IV. Conclusion

Thus, based on experiments with cloud infrastructure and different database decompositions, following features were observed:

-The need to use specialized tools to control feedback on routing and protocol levels.

-It is required, starting with development phase of the software application, to include the possibility of changing the database structure.

References

[1] M. Bahrami, M. Singhal, “The role of cloud computing architecture in big data,” Information Granularity, Big Data, and Computational Intelligence, Vol. 8, pp. 275-295, 2015.

[2] Lackermair G. “Hybrid cloud architectures for the online commerce”. Procedia Computer Science, vol. 3, pp. 550-555, 2011. doi:10.1016/j.procs.2010.12.091

[3] Javadi, B., Abawajy, J., & Buyya, R., “Failure-aware resource provisioning for hybrid cloud infrastructure,” Journal of parallel and distributed computing, vol. 72, pp. 1318-1331, 2012.

[4] Kaviani N., Wohlstadter E., Lea R., “MANTICORE: A framework for partitioning software services for hybrid cloud,” 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom), pp. 333-340, 2012. doi: 10.1109/CloudCom.2012.6427541

[5] Sotomayor, B., Montero, R. S., Llorente, I. M., & Foster, I., “Virtual infrastructure management in private and hybrid clouds,” IEEE Internet computing, vol. 13, pp. 14-22, 2009. doi: 10.1109/MIC.2009.119

[6] Naik, P., Agrawal, S., & Murthy, S., “A survey on various task scheduling algorithms toward load balancing in public cloud,” American Journal of Applied Mathematics, vol. 3, pp. 14-17, 2015.

[7] Fowley, F., Pahl, C., & Zhang, L., “A comparison framework and review of service brokerage solutions for cloud architectures,” Service-Oriented Computing–ICSOC 2013 Workshops, pp. 137-149, 2014. doi: 10.1007/978-3-319-06859-6_13

[8] Pluzhnik, E., Nikulchev, E., “Virtual laboratories in cloud infrastructure of educational institutions,” Proceedings 2nd International Conference on Emission Electronics (ICEE) Selected papers, 2014. doi: 10.1109/ICCTPEA.2014.6893324

[9] Pluzhnik E., Nikulchev E., Payain S., “Optimal control of applications for hybrid cloud services,”. Proceedings 2014 IEEE 10 World Congress on Services, pp. 458-461, 2014. doi: 10.1109/SERVICES.2014.88

[10] Lukyanchikov O., Payain S., Nikulchev E., Pluzhnik E. “Using object-relational mapping to create the distributed databases in a hybrid cloud infrastructure”. International Journal of Advanced Computer Science and Applications, vol. 5, no. 12, pp. 61-64, 2014. doi: 10.14569/IJACSA.2014.051208

Intelligent Wireless Indoor Monitoring System based on ARM

Jia Chunying
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

Zhang Liping*
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

Chen Yuchen
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

***Abstract*—This paper proposed an intelligent wireless indoor monitoring system based on STM32F103. The system compromises a master and terminals, which communicates through a CC1101 433M wireless unit. Using ENC28J60 and SIM900A to access Ethernet, the system can intelligent push information to cloud server, therefore we can observe and control terminals remotely. This paper present an algorithm based on MCU ID for terminal code recognition. The algorithm can intelligently discriminate terminals through code matching code between master and terminals, though hardware and software of terminals is similar. The experiment demonstrates that the proposed system is robust stability, real-time and obtains accurate warning information.**

Keywords—STM32; CC1101; Intelligent Monitoring; ID; Corresponding Code Discrimination

I. Introduction

With the rapid development of China's network technology, information technology/economy, people's requirements of the life quality increase gradually. In such a high-paced living environment, traditional indoor video surveillance systems appear to be inadequate. Intelligent interior monitoring system uses the cloud server, which can real-time detect indoor state when an exception occurs, automatically alarm and timely send the warning to user. Therefore, it can greatly reduce the family security risks.

In this paper, an intelligent system based on STM32F103 is designed to predict household hazards, and send the indoor gas concentrations and smoke density information, situation of indoor power, emergency data to the monitoring center through both wired and wireless network.

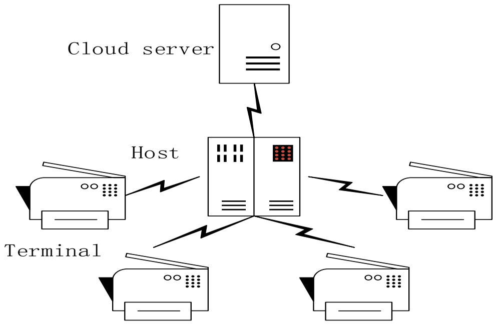


Fig. 1. Monitoring system proposed

As shown in Figure 1, the system consists of a host and terminals, which communicates each other through 433M wireless units.

II. Monitoring Host

As shown in Figure 2, the Monitoring host contains a 433M wireless communication unit, a WIFI and a RFID module, an Ethernet and a GPRS / GSM module.

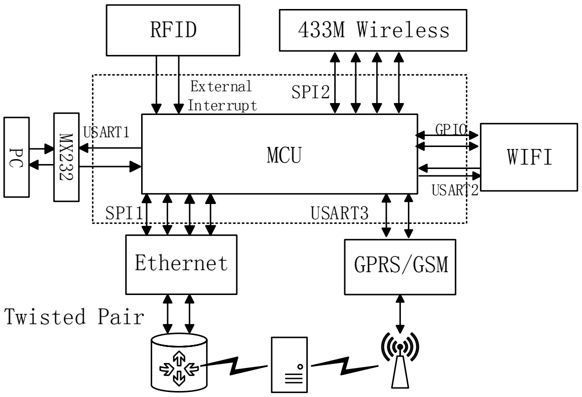


Fig. 2. monitoring host

Host Communicate with the terminal through a 433M wireless module. Normally the host uses a query method to get terminal operating status. The monitoring terminal sends alarm information to the host immediately if an abnormal condition is detected. The host sends the information to the cloud server through wired Ethernet or GPRS / GSM. The server pushes messages to user's phone, to achieve the purpose of real-time monitoring.

III. Hardware

A. *Cortex-M3 STM32F103VET6*

We used a STM32F103VET6 MCU as the system's core, which is produced by ST Microelectronics. STM32F103VET6 is based on ultra-low-power ARM Cortex-M3 processor core. The STM32F103VET6 series are a powerful lineup in 32 bit STM32 microcontroller family. They have good compatibility. ARM Cortex-M3 core set is specialized in high-efficiency low-energy in real-time applications. It has high cost-effective which embedded developers favor so much. The enhanced chip with 512KB RAM and 64KB SRAM has the maximum operating frequency of72MHz, up 1.25DMips / MHz [1] at zero wait cycle memory access.

The chip has more than three SPI (serial peripheral interface). The full duplex or half-duplex communication rate of SPI can be up to 18 Mbit /sec in slave or master mode. It has a 3 bit prescaler which produces eight kinds of master mode. Its frequencies can be configured as 8 or 16 bits per frame. Hardware CRC generation / verification supports basic SD card and MMC mode. All SPI interfaces can use DMA operations. It has 3 UART and 2 UART serial communication interface, USART1 interface communication rate up to 4.5 Mbit / sec interface speed up other communication 2.25 Mbit / s[1].

B. 433M wireless communication module

1) Introduction and Application CC1101

Host and terminal use a CC1101 chip to realize communication. CC1101 is a high-performance sub-1GHz wireless transceiver designed for low-power RF applications. It's mainly used for industrial, scientific and medical (ISM) and short-range device (SRD). CC1101 provides packet handling, data buffering, burst transmissions, received signal strength indication (RSSI), clear channel assessment (CCA), Wireless on wake (WOR), extensive hardware support. CC1101 and CC1100 are compatible in code, package and pin-out. It is open to sub-1GHz frequency RF design which is commonly used in world [2]. Interfaces between MCU and CC1101 are shown in Figure 3.

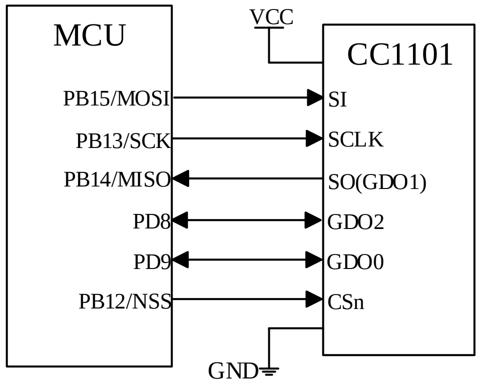


Fig. 3. MCU and CC1101

2) Wireless module

CC1101 have five work modes, part of the idle mode, sleep mode, receive mode, sent mode and crystal oscillator close mode. The work mode of the chip can be change by the internal set register of the external pin. At first electrify, the chip will enter the sleep mode, pause work, but the value of internal chip register can't be lost. When wireless module is in a dormant state, you can pin CSn is pulled low, so that the module enters an idle state. You can read and write the chip registers and mode settings when the wireless module enters an idle state. When we set inside STX bit register, crystal is startup, and CC1101 wireless module enters a transmission mode. If the bit is set SRX, the wireless module of CC1101 enters into the receiving state [3].

When they need to send data, STM32F103ZET6 writes data to wireless module at a lower rate through SPI. RF circuits start after writing completed while RF circuits send written data out and empty TX-FIFO register. When wireless module received data, it sent the received data is to STM32F103ZET6 through the SPI. Because the wireless module can enter sleep mode to reduce power consumption, CC1101 wireless module can automatically generate the CRC and preamble code, which reduces the complexity of system design.

3) Ethernet Module

We used ENC28J60 for Ethernet communication. ENC28J60 is the first 28-pin stand-alone Ethernet controller over the world. It is produced by Microchip and provides a low pin count, low-cost, streamlined remote communications solutions for embedded systems. ENC28J60 is an independent Ethernet controller with an industry-standard serial peripheral interface. It meets all IEEE 802.3 specification, using a series of packet filtering schemes to limit incoming packets. It also provides an internal DMA module for fast data throughput and hardware support for IP checksum calculation. Communication with the host controller is provided by two interrupt pin (INT and WOL) and SPI pins (SO, SI, SCK, CS). The data transfer rates of ENC28J60 is up to 10Mb / s. Two dedicated pins (LEDA, LEDB) is used for connecting a LED which indicators the network activity. Figure 4 is the wiring diagram for the MCU with ENC28J60.

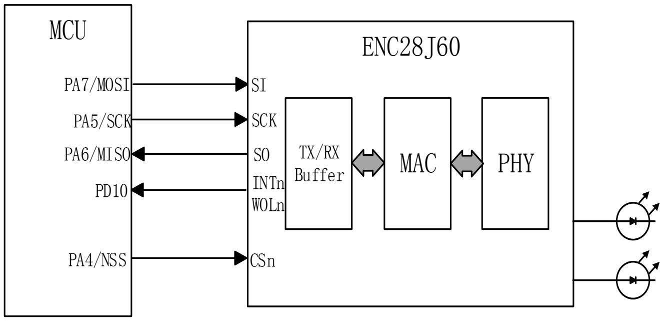


Fig. 4. ENC28J60 and MCU circuit

IV. Software System

A. The host software and system structure

In this system, we used Keil uVision5 MDK as development tool. It was officially released by ARM on October 2013. It includes data collection procedures Wiegand ID card, SPI driver, and TCP / IP communication protocol LwIP protocol, SIM900A AT command communication program, serial communication program and the main loop procedures. Master program flow is shown in Figure 5.

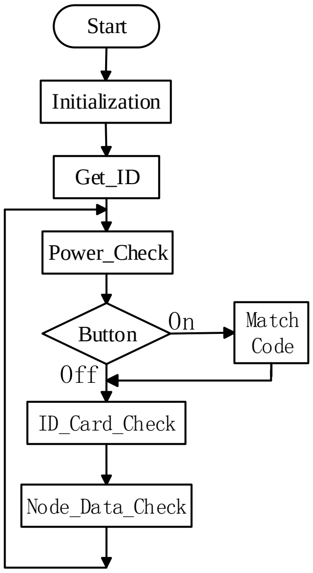


Fig. 5. Figure 5 flowchart

The main program is mainly responsible for components initialization. In software systems, Because of the control parts is fewer, we do not use the operating system. The matching code switch is used to discriminate the same set of identification for communication between devices.

B. Communication between the host and terminal

1) The Communication Protocol

In this design, the communication between host and terminal uses a 433M wireless module, which has strength signal, high anti-jamming etc. The communication included request, response, and transmission, closed four states. The switch of communicate state is shown in Figure 6.

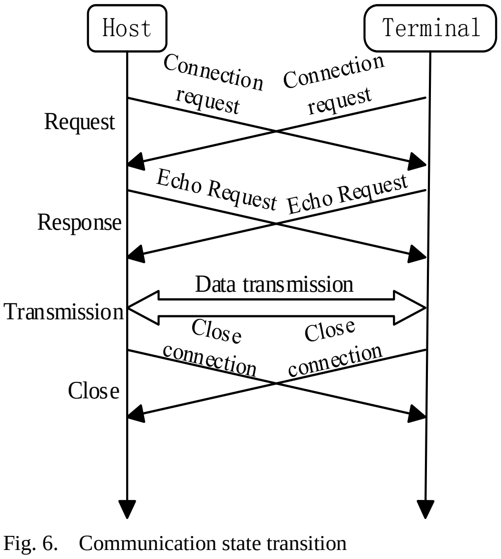


Fig. 6. Communication state transition

2) Encryption Algorithm

In order to ensure secure communications, the design uses a lightweight ECC encryption algorithm. Elliptic Curves Cryptography (ECC) is a public key encryption system, which is proposed in 1985 by Koblitz and Miller. The mathematical basis of ECC is rational points on an elliptic curve additive group constituting the Abel elliptic discrete logarithm computational difficulty.

Compared with classic RSA, DSA, and other public key cryptosystem, elliptic cryptosystem has high security, processing speed, small storage space, and low bandwidth requirements. Some studies indicate that elliptical key 160 and 1024 has the same RSA key security. On the private key encryption and decryption, ECC algorithm has a faster speed than RSA, DSA. Elliptic Curves Cryptography flow is shown in Figure 7.

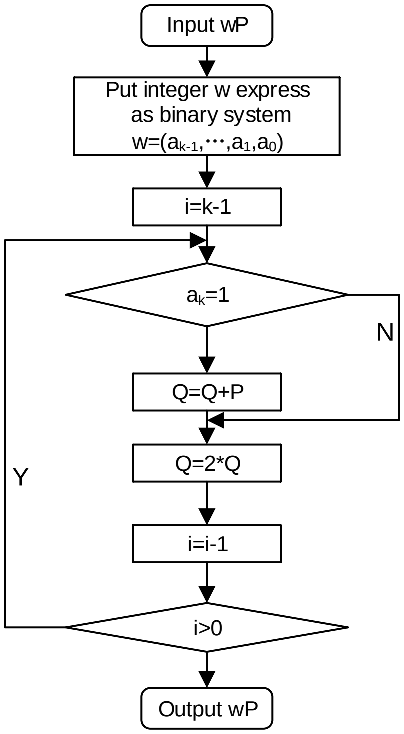

Input wP
Put integer w express as binary system w=(a_{k-1},⋯,a_1,a_0)
i=k-1
a_k=1
N
Q=Q+P
Q=2*Q
Y
i=i-1
i>0
Output wP

Fig. 7. Flow chart of Elliptic Curves Cryptography

3) The match code identification solutions based on MCU ID

Since the terminal can freely expand, there will have an identical terminal device. The host cannot identify the terminal which sends information it received. To solve the problem, we propose a solution to identifying terminal by MCU ID Code.

a) Electronic Signature

Electronic signature stored in the system memory area of the flash memory module, you can use JTAG/SWD or CPU to read. Chip identification information contains writing at the factory. User firmware or external device can read an electronic signature, to automatically match the different configurations SRM32F10xxx microcontroller.

b) MCU 96 bit unique identity ID

A 96 bit of unique identity is provided to any one of the STM32 microcontroller. Users under any circumstances cannot change this identity. The 96 bit unique identity of the product, according to the usage of different users, can be read in bytes

(8 bit) as a unit, or can be read in half-word (16 bit) or whole-word (32 bit). Product unique identity is very suitable to be used as 1) A sequence number (such as USB serial numbers or other characters in the terminal application) 2) A password in the preparation of flash memory; 3) In conjunction with this uniquely identifies software encryption and decryption algorithms to improve code safety in the flash memory; 4) Safety mechanism to activate the bootstrap process. The structure of the ID storage is shown in Figure 8.

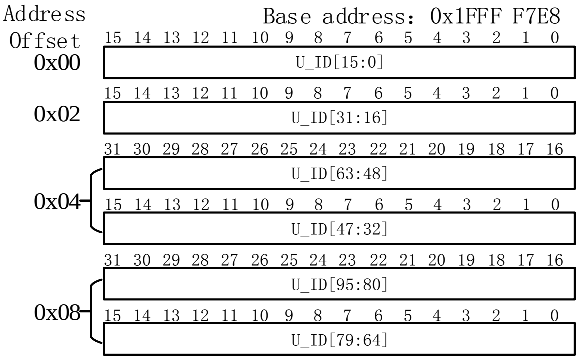


Fig. 8. ID Code Storage Structure

c) Reads and encrypted ID code

Use the following code in the main program can read the MCU 96 ID.

```
CpuID[0]=*(vu32*)(0x1ffff7e8);
CpuID[1]=*(vu32*)(0x1ffff7ec);
CpuID[2]=*(vu32*)(0x1ffff7f0);
```

Considering the practical use does not need to 96 bit, as well as security using simple encryption algorithm compresses 96 to 24-bit codes, the encryption algorithm is;

```
ID_Code=(CpuID[0]>>1)+(CpuID[1]>>2)+(CpuID[2]>>3;
```

d) Matching Code

In this paper, for security reasons, we uses a match code switch in hardware. The hardware circuit is shown in Figure 9.

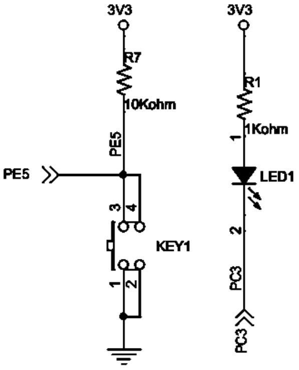


Fig. 9. code switching circuit

KEY1 is the match code button. When the button is pressed for more than 2s, it will start the program of match code. Master sends and saves ID Code, while the success of matching code is returned. As the matching code indicator light, LED1 will flash 5 times in frequency of twice per second after success of matching code.

V. Experimental Results and Analysis

We selected a 4 rooms and 2 hall area of 128 square meters for experimental. A host was placed in a drawing room, and terminals were placed in cook room (gas monitor), bathroom (emergency button), study room (smog monitoring), master bedroom (bed vacancy monitor). Test results are shown in Table 1 and Table 2.

TABLE I. Ethernet is Connected

	Monitoring host Response delay	Server response delay
Cook room	326ms	3.2s
Bathroom	216ms	2.1s
Study room	302ms	1.3s
Bedroom	416ms	3.2s

TABLE II. Ethernet is Disconnected

	Monitoring host Response delay	Server response delay
Cook room	413ms	10.6s
Bathroom	321ms	11.3s
Study room	215ms	10.7s
Bedroom	258ms	12.5s

As shown in Table 1 and Table 2, it can be seen that he server response time is about 2.5s when Ethernet is connected, whereas the server response time is about 12s when the Ethernet is disconnected. Because this phenomenon is that SIM900A uses GPRS to communicate while GPRS transmission rate is only about 56 to 114Kbps . However, 10Mbps is a certain gap for ENC28J60. System design taking into consideration factors such as cost and stable areas, there is no choice of the current 3G/4G communications technology.

VI. Conclusion

The experiments show that the control system design is reasonable, reliable operation and has good scalability comparing to traditional video surveillance system. The alarm information is accurate, targeted and real-time. The system pushes cloud exception information to cloud server which is used remotely observe and control terminals.

Acknowledgment

First and foremost, I would like to show my deepest gratitude to my supervisor, Prof. Zhang Liping, a respectable, responsible and resourceful scholar, who has provided me with valuable guidance in every stage of the writing of this paper. Without her enlightening instruction, impressive kindness and patience, I could not have completed my paper. Her keen and vigorous academic observation enlightens me not only in this paper but also in my future study. I would also like to thanks my teacher Prof. Chen Yuchen he have helped me to develop the fundamental and essential academic competence.

References

[1] STMicroelectronics.STM32Reference annual (RM0008). http:// www. stm. com . 2012, 16-18,69,70,422,472.

[2] Texas Instruments.CC1101 user's manual[EB/OL].2008.http://focus.ti. Com/lit/ds/symlink1/c 101. pdf.

[3] Nan Yimin. Based STM32F103xxx peripheral device programming STM32 standard peripheral libraries [J]. Journal of Changsha Aeronautical Vocational and Technical College. 2010(04),41-45

[4] CC1101 Single－Chip Low Cost Low Power RF－Transceiver, Data Sheet [Z],Texas Instruments,2008.

[5] WIFI chip of low-power specification ESP8266.[s] 2013.12 5,6

[6] Hosam El-Ocla. TCP CERL: congestion control enhancement over wireless networks[J]. Wireless Networks , 2010, 16(1):183-198

[7] Wang Mingxin. GSM Remote Monitoring System Based on SIM900A[J]. Journal of computer knowledge and technology. 2014(15):3500-3504.

[8] Wang Lei, Wang Jun. Embedded remote appliance controller design Based on SIM900A[J].2014,27(1):76-80.

[9] lwIP--A Lightweight TCP/IP stack[EB / OL]. http://savannah.nongnu. org /projects/lwip/ .

[10] Han Zexi, Lv Fei. Research and development of GSM network AT command emulation system [J]. Modern electronic technology. 2005,28(17):9-11.

[11] Bi Dandan, Li Yun, Feng Zhi. Smart car design based on MCU and GSM modules [J]. Data communication. 2014,(04):41-43

[12] Wei Hongtao, Wong Huijuan, Wu Xixiu. The Design and implementation of emergency communication control system Based on 3G[J]. Journal of Wuhan University of Technology Information & Management Engineering. 2014,36(04):443-446.

[13] Que Fanbo. The Design of program Remote upgrade Based on STM32. [J]. Electronic Instrumentation Customer. 2013(05)，90-92

Markovian Process and Novel Secure Algorithm for Big Data in Two-Hop Wireless Networks

K. Thiagarajan,
Department of Mathematics,
PSNA College of Engineering and Technology,
Dindigul, India.

A. Veeraiah,
Department of Mathematics,
K. L. N College Engineering and Technology,
Madurai, India.

K. Saranya,
Department of Information Technology,
PSNA College of Engineering and Technology,
Dindigul, India.

B. Sudha,
Department of Mathematics
SRM University,
Chennai, India.

***Abstract*—This paper checks the correctness of our novel algorithm for secure, reliable and flexible transmission of big data in two-hop wireless networks using cooperative jamming scheme of attacker location unknown through Markovian process. Big data has to transmit in two-hop from source-to-relay and relay-to-destination node by deploying security in physical layer. Based on our novel algorithm, the nodes of the network can be identifiable, the probability value of the data absorbing nodes namely capture node C, non-capture node NC, eavesdropper node E, in each level depends upon the present level to the next level, the probability of transition between two nodes is same at all times in a given time slot to ensure more secure transmission of big data. In this paper, maximum probability for capture nodes is considered to justify the efficient transmission of big data through Markovian process.**

Keywords*—big data; two-hop transmission; security in physical layer; cooperative jamming; energy balance; Markov process

I. Introduction

Wireless networks have become an indispensable part of our daily life, used in many applications where the amount of data is very massive and is called big data [2]. Security is a critical issue in wireless applications of big data, when people rely heavily on wireless networks for transmission of important/private information. Therefore, the ability to share secret information reliably in the presence of eavesdroppers is extremely important in the environment of big data [8]. Cryptographically enforced security is not sufficient to provide everlasting security in handling huge data size due to increased attacks by capturing its keys [10]. So security in physical layer is used to retain the everlasting security in big data as it prevents eavesdroppers and malicious nodes from capturing the data [11].

The term big data is high volume, variety, velocity and veracity. The amount of data increases faster and quicker in big data. According to a report published by IBM in 2012 [4], 90 percent of the data in the world was generated in the previous two years. As a consequence, the concept of the big data has emerged as a widely recognized trend, which is currently attracting much attention from government, industry, and academia [3]. It is essential to have data transfer mechanism, two-hop transmission from source-to-relay and relay-to-destination node plays a vital role in secure and energy efficient transmission of big data.

Cooperative communication helps in exploiting spatial diversity to enhance the quality of wireless links. The characteristics of cooperative networks are shown in Fig. 1. Security can be improved by cooperative networks by having the information content minimum to the eavesdropper nodes of the expected destination and having maximum to the relay node of the expected destination [6]. The recently proposed cooperative network technique is cooperative jamming to improve physical layer security in the presence of eavesdroppers [5]. In wireless communication, occurrence of interference is considered redundant. This fetches the work of cooperative jamming for flexible and efficient wireless network technique to confuse the eavesdroppers and making the source message uncertain by generating friendly jamming signal to the eavesdroppers. In this, if the data has to be transmitted from source S to destination D, jamming signal will be emitted by the relay nodes to have the secure communication and to prevent the eavesdroppers of location

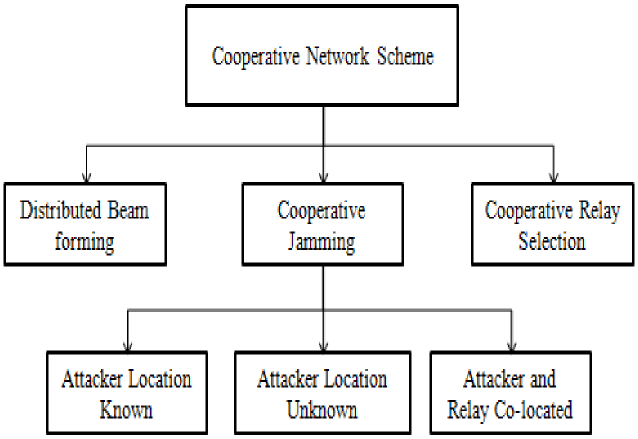

Cooperative Network Scheme
Distributed Beam forming
Cooperative Jamming
Cooperative Relay Selection
Attacker Location Known
Attacker Location Unknown
Attacker and Relay Co-located

Fig. 1. Classification of cooperative network scheme

unknown from capturing the data. In our novel algorithm, cooperative jamming scheme is considered [9].

The remainder of this paper is organized as follows. Section II highlights the overview of novel secure algorithm [9]. Section III discuss about the Markovian implementation on the proposed algorithm. We conclude our paper by experimental result verification for the proposed algorithm.

II. Overview of Novel Secure Algorithm

The Fig. 2 shows the overview diagram for the novel secure transmission algorithm which is clearly explained in our previous paper [9], in which we want to select the data transmitting region which is of side length l followed by segmenting the selected region of equal size. Then we have to determine the probability value for each node which is detailed in our paper [9]. Based on probability value for capture node we want to classify the transmission as secure data transmission and unsecure data transmission. If it is secure data, transmit the data in two-hop. If it is unsecure data transmission, transmit the data in two-hop by adopting cooperative jamming technique to prevent eavesdroppers from capturing the data. The jammers should be of distance r away from the intended destination. We assume that only one end-to-end transmission can be conducted in one time slot.

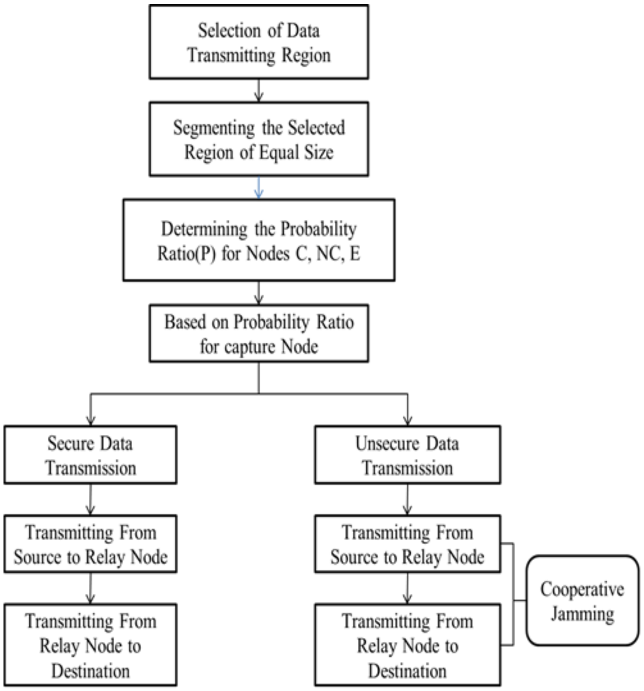

Selection of Data Transmitting Region
Segmenting the Selected Region of Equal Size
Determining the Probability Ratio(P) for Nodes C, NC, E
Based on Probability Ratio for capture Node
Secure Data Transmission
Unsecure Data Transmission
Transmitting From Source to Relay Node
Transmitting From Relay Node to Destination
Transmitting From Source to Relay Node
Transmitting From Relay Node to Destination
Cooperative Jamming

Fig. 2. Overview of proposed algorithm

III. Discussion of Markovian Process on Proposed Algorithm

A. Markov Model

Markovian model is a model of representing different resident states of a system, and the transitions between the different states [7]. Similarly in the algorithm which is proposed in [9] have different states of the system namely source, capture, non-capture, eavesdropper, ideal and destination nodes for having transition between different states to have secure transmission of big data.

B. Stochastic Process

Let S be a sample space of a stochastic experiment. A stochastic process is a mapping X which assigns to every outcome s ϵ S a real valued function of time x (t, s) (i.e) X (s) = x (t, s). The family or ensemble of all such time functions is denoted by X (t, s) and is called a stochastic process [7]. The novel secure transmission algorithm which we have proposed in [9] satisfies with the stochastic process in which the behavior of the system varies randomly with time and space for each end-to-end transmission of big data.

C. Markov Process

Markov process is the simplest generalization of independent processes which allow the outcome at any instant to depend only on the outcome that precedes it and not on the earlier ones [7]. As per our algorithm, a stochastic process X(t) is said to be a Markov process if for any $t_1 < t_2 < t_3 < \ldots\ldots < t_n$

$$P[X(t_n) \le x_n \mid X(t_{n-1})$$

$$= x_{n-1}, X(t_{n-2}) = x_{n-2}, \ldots\ldots X(t_1) = x_1]$$

$$= P\,[X(t_n) \le x_n \mid X(t_{n-1}) = x_{n-1}]$$

(i.e.) the conditional distribution of X (t_n) for given values of X (t_1), X (t_2), …,X(t_{n-1}) depends only on present state X(t_{n-1}).

D. Markov Chains

Based on our algorithm in [1], let X (t) be a Markov process with states X (t_r) = X_r , $t_0 < t_1 < \ldots.. < t_n$. If for all n,

$$P\,[X_n = a_n \mid X_{n-1} = a_{n-1}, X_{n-2} = a_{n-2}. \ldots., X_0 = a_0]$$

$$= P\,\{X_n = a_n \mid X_{n-1} = a_{n-1}]$$

then the sequence of random variables {Xn} is called a Markov chain, n=0,1,2… Here $a_{1,}\ a_2 \ldots a_n$ are called the states of Markov chain [7].

E. Transition Probabilities

P {$X_m = a_i$} = P_i (m) represents the probability that at time t=t_m, the system occupies the state a_i, P [$X_n = a_j \mid X_m = a_i$] = P_{ij}(m, n) represent the probability that system goes into state a_j at t=t_n given that it was in state ai at t=t_m. The numbers P_{ij}(m,n) represent the transition probabilities of the Markov chain from state a_i at time tm to a_j at time t_n [7].

F. Markov Transition Diagram

The state of transmitting the data transition probabilities from i^th^ state to the j^th^ state by an arc labeled as P_{ij} is the representation of Markov state transition diagram is shown in Fig. 3.

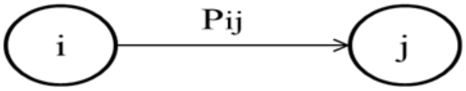


Fig. 3. Model transition diagram

G. Markov Transition Matrix

$$P = \begin{array}{cc} & \begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \cdots \\ m \end{array} & \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ P_{m1} & P_{m2} & \cdots & P_{mn} \end{bmatrix} \end{array}$$

The square matrix represents the number of states in rows and columns [7]. The rows are represents the nodes from which the data is transmitting. The columns are representing the nodes to which the data is received. The sum of the probability in each row must equals to 1.

IV. Experimental Result Verification

Discrete time and space Markov chain is used to verify our novel secure transmission algorithm. As in our previous work [9] and with the reference from [1, 12] secure transmission of big data through binary probability evaluation value in Table I. Based on that result we are going to verify using Markov process.

TABLE I. Binary Evaluation to Verify the Transmission

C	NC	E	Transmission
0	0	0	No Action
0	0	1	Unsecure
0	1	0	Unsecure
0	1	1	Unsecure
1	0	0	Secure
1	0	1	Secure/Unsecure
1	1	0	Secure/Unsecure
1	1	1	Secure/Unsecure

The verification process is based on minimum probability is for the non-capture and eavesdropper nodes. The maximum probability is for the capture node to prove the secure transmission of big data. If the probability is Maximum for the malicious nodes it can be prevented to capture the big data by cooperative jamming [9].

A. Case I

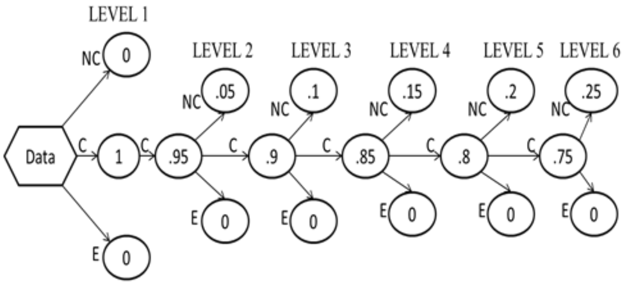
With eavesdroppers E=0, sum of the values for capture node C, non-capture node NC is 1. The verification result to this case is discussed in Markov Principle as transition table in Table II, transition diagram in Fig. 4 and by transition matrix.

a) Probability Values Distribution Table

TABLE II. Transition Table (Case I)

C	NC	E
1	0	0
0.95	0.05	0
0.9	0.1	0
0.85	0.15	0
0.8	0.2	0
0.75	0.25	0
……………………………………N-1, N		

b) Markov Transition Diagram



……………… LEVEL N-1, LEVEL N

Fig. 4. Transition diagram (case I)

c) Markov Transition Matrix

$$\text{LEVEL 1} \quad \begin{array}{cc} & \begin{array}{ccc} C & NC & E \end{array} \\ \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \text{LEVEL2} \quad \begin{array}{cc} & \begin{array}{ccc} C & NC & E \end{array} \\ \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 0.95 & 0.05 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \text{LEVEL 3} \quad \begin{array}{cc} & \begin{array}{ccc} C & NC & E \end{array} \\ \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$\text{LEVEL 4} \quad \begin{array}{cc} & \begin{array}{ccc} C & NC & E \end{array} \\ \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 0.85 & 0.15 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \text{LEVEL 5} \quad \begin{array}{cc} & \begin{array}{ccc} C & NC & E \end{array} \\ \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \text{LEVEL 6} \quad \begin{array}{cc} & \begin{array}{ccc} C & NC & E \end{array} \\ \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}$$

…………………LEVEL N-1, LEVEL N

⇨ COMPLETE MARKOV CHAIN

$$\begin{array}{cc} \begin{array}{c} C \\ NC \\ E \end{array} & \begin{bmatrix} 1 & 0.95 & 0.9 & 0.85 & 0.8 & 0.75 & \cdots & N-1 & N \\ 0 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & \cdots & N-1 & N \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & N-1 & N \end{bmatrix} \end{array}$$

Since the eavesdropper probability value is zero in the above case, obviously the big data transmission is more secured and the novel algorithm [9] satisfies with this case.

B. Case II

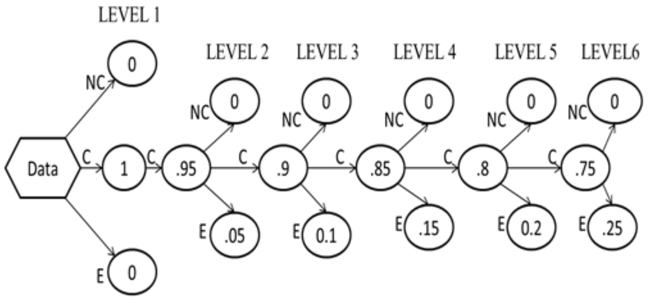
With non-capture node NC=0, and summation of capture and eavesdropper node is 1 (i.e., $\sum (C_i + E_i = 1)$. This case is verified in the following Table III as transition table, in Fig. 5 as transition diagram as well as in transition matrix of Markov principle.

a) Probability Value Distribution Table

TABLE III. Transition Table (Case II)

C	NC	E
1	0	0
0.95	0	0.05
0.9	0	0.1
0.85	0	0.15
0.8	0	0.2
0.75	0	0.25
………………………………..N-1, N		

b) Markov Transition Diagram



……………… LEVEL N-1, LEVEL N

Fig. 5. Transition diagram (case II)

c) Markov Transition Matrix

LEVEL 1

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 1 & 0 & 0 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL2

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.95 & 0 & 0.05 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.9 & 0 & 0.1 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 4

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.85 & 0 & 0.15 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 5

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.8 & 0 & 0.2 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 6

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.75 & 0 & 0.25 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

…………………LEVEL N-1, LEVEL N

⇨ COMPLETE MARKOV CHAIN

$$\begin{array}{c} \mathrm{C} \\ \mathrm{NC} \\ \mathrm{E} \end{array} \begin{bmatrix} 1 & 0.95 & 0.9 & 0.85 & 0.8 & 0.75 & \cdots & \mathrm{N}-1 & \mathrm{N} \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \mathrm{N}-1 & \mathrm{N} \\ 0 & 0.05 & 0.1 & 0.15 & 0.2 & 0.25 & \cdots & \mathrm{N}-1 & \mathrm{N} \end{bmatrix}$$

The above case with non-capture nodes probability value as zero identifies that the transmission of big data is more secured by setting the probability for eavesdropper node as minimum. If the eavesdroppers try to capture the data it will be prevented by cooperative jamming scheme [9].

C. Case III

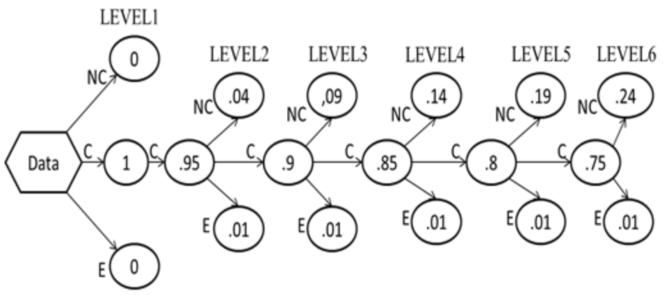
With addition of capture node C, non-capture node NC, and eavesdropper node E value as 1(i.e., C_i + NC_i +E_i =1). The Table IV and Fig. 6 check the correctness of Markov rule to this case in our proposed algorithm.

a) Probability Value Distribution Table

TABLE IV. TRANSITION TABLE (CASE III)

C	NC	E
1	0	0
0.95	0.04	0.01
0.9	0.09	0.01
0.85	0.14	0.01
0.8	0.19	0.01
0.75	0.24	0.01
……………………………………..N-1, N		

b) Markov Transition Diagram



…………………LEVEL N-1, LEVEL N

Fig. 6. Transition diagram (case III)

c) Markov Transition Matrix

LEVEL 1

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 1 & 0 & 0 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL2

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.95 & 0.04 & 0.01 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 3

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.9 & 0.9 & 0.01 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 4

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.85 & 0.14 & 0.01 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 5

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.8 & 0.19 & 0.01 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

LEVEL 6

$$\begin{array}{cccc} & \mathrm{C} & \mathrm{NC} & \mathrm{E} \\ \mathrm{C} & 0.75 & 0.24 & 0.01 \\ \mathrm{NC} & 0 & 0 & 0 \\ \mathrm{E} & 0 & 0 & 0 \end{array}$$

…………………LEVEL N-1, LEVEL N

⇨ COMPLETE MARKOV CHAIN

$$\begin{array}{c} C \\ NC \\ E \end{array} \begin{bmatrix} 1 & 0.95 & 0.9 & 0.85 & 0.8 & 0.75 & \cdots & N-1 & N \\ 0 & 0.04 & 0.09 & 0.14 & 0.19 & 0.24 & \cdots & N-1 & N \\ 0 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & \cdots & N-1 & N \end{bmatrix}$$

The above Markov process haves probability for all naming nodes ensures secured transmission and proves for proposed novel secure transmission algorithm [9].

V. CONCLUSION AND FUTURE WORK

The verification through Markovian process revealed that proposed novel secure transmission algorithm is more secure and energy efficient to transmit big data by our binary based evaluation process with minimum probability at eavesdroppers and maximum at capture node. In future the proposed algorithm will be discussed and verified through finite state automaton (FSA).

ACKNOWLEDGEMENT

The authors would like to thank Dr. Ponnammal Natarajan worked as Director–Research, Anna University-Chennai, India and Dr. K. Sarukesi former vice chancellor in Hindustan University-Chennai, India for their cognitive ideas and dynamic discussions with respect to the paper's contribution.

REFERENCE

[1] Almudena Konrad, Ben Y. Zhao, Anthony D. Joseph, Reiner Ludwig "A Markov-Based Channel Model Algorithm for Wireless Networks", http://bnrg.cs.berkeley.edu/~adj/publications/paper-files/winet01.pdf.

[2] Chip Craig J. Mathias Principal, Farpoint Group COMNET 2003 —"Wireless Security: Critical Issues and Solutions" 29 January 2003.

[3] A. Divyakant, B. Philip, and et al., "Challenges and opportunities with Big Data," 2012, a community white paper developed by leading researchers.

[4] IBM, "Four vendor views on big Data and big data analytics: IBM," http://www-01.ibm.com/software/in/data/bigdata/, Jan. 2012.

[5] R. Negi, S. Goel, "Secret communication using artificial noise", in: IEEE Vehicular Technology Conference, 2005, pp. 1906–191.

[6] P. Popovski and O. Simeone, "Wireless secrecy in cellular systems with infrastructure-aided cooperation", IEEE Trans. Inf. Forensics Security,vol. 4, no. 2, pp. 242–256, Jun. 2009.

[7] S. M. Ross, "Stochastic Processes", John Wiley and Sons, 1996.

[8] R. Schell, "Security – a big question for big data", in: IEEE International Conference on Big Data, pp. 5–5, October 2013.

[9] K. Thiagarajan, K. Saranya, A. Veeraiah, B Sudha, "Wireless Transmission of Big Data Using Novel Secure Algorithm", in 17th International conference on Mathematical Sciences, Engineering and Application. WASET- June 2015., in Press.

[10] J. Talbot, D. Welsh, "Complexity and Cryptography: An Introduction", Cambridge University Press, 2006.

[11] A. D. Wyner, "The wire-tap channel", Bell Syst. Tech. J. 54 (8) (1975) 1355–1387.

[12] M. Zorzi., and R. R. Rao.," On the statistics of block errors in bursty channels". In IEEE Transactions on Communications (1997).

Author Profile

Dr. K.Thiagarajan working as Associate Professor in the Department of Mathematics in PSNA College of Engineering and Technology, Dindigul, Tamil Nadu, India. He obtained his Doctorate in Mathematics from University of Mysore, Mysore, India in Feb 2011. He has totally 14 years of experience in teaching. He has attended and presented 38 research articles in national and international conferences and published one national and 42 international journals. Currently he is working on web mining and big data analytics through automata and set theory. His area of specialization is coloring of graphs and DNA computing in Ph.D. program.

K. Saranya received B.E in Computer Science and Engineering from PSNA College of Engineering and Technology, Dindigul, affiliated to Anna University-Chennai, India in 2013. Currently she is pursuing masters in Computer Science and Engineering (with Specialized in Networks) in PSNA College of Engineering and Technology, Dindigul, affiliated to Anna University-Chennai, India.She presented papers in conferences and published in international journals. Currently she is working on big data analytics.

A. Veeraiah completed M.Sc, M.Phil Madurai Kamaraj University (School Of Mathematics) Madurai. He is a Gold medalist of M.sc in Mathematics. Currently he is working as Associate Professor in K.L.N College of Engineering, Pottapalayam, Tamil Nadu, India. He has totally more than 10 years of teaching experience in UG and PG Level. He has passed SET exam conducted by Bharathiyar University, Coimbatore, during the year October 2012

B. Sudha received M.Phil degree at Bharathidhasan University, Trichy, India in 2010. Currently she is working as a Assistant Professor in SRM university, Chennai, India. She is a life member of Indian mathematical society (IMS). She also presented papers in conferences and published in international journals.

Towards GP Sentence Parsing of V+P+CP/NP Structure

A Perspective of Computational Linguistics

Du Jiali
School of Foreign Studies
Nanjing University
Nanjing, China

Yu Pingfang
Faculty of Chinese Language and Culture
Guangdong University of Foreign Studies
Guangzhou, China

***Abstract*—Computational linguistics can provide an effective perspective to explain the partial ambiguity during machine translation. The structure of V+Pron+CP/NP has the ambiguous potential to bring Garden Path effect. If Tell+Pron+NP structure has considerable higher observed frequencies than Tell+Pron+CP structure, the former is regarded as the preferred structure and has much lower confusion quotient. It is possible for the grammatical unpreferred Tell+Pron+CP structure to replace the ungrammatical preferred Tell+Pron+NP, which results in the processing breakdown. The syntactic details of GP processing can be presented by the computational technologies. Computational linguistics is proved to be effective to explore the Garden Path phenomenon.**

Keywords—artificial intelligence; computational linguistics; machine learning; local ambiguity; garden path sentences

I. INTRODUCTION

A Garden Path (GP) sentence can produce an effect of processing breakdown. It refers to the special local ambiguous structure which always makes a reader misled down to the garden path. The GP sentence is grammatically correct and the ultimate result of processing is different from the initial incorrect interpretation which is considered to be the most likely one at first. The ups and downs of processing may lure the reader into a dead one. With the appearance of garden path effect, the initially built-up structure is replaced with the improved structure whose processing is involved in the breakdown and backtracking. Connectionist psycholinguistics focus on psychology of language, and connectionist models are considered to play an important role in GP processing.[1]

The parsing of GP sentences is complex during the processing of natural language. Information-based decision-making provides a helpful information method to analyze the complex structure, i.e. GP sentence, by describing the phenomenon of information use and by explaining why an information use occurs as it does. The relative effectiveness can be obtained by using multi-term phrases and POS tagged terms. [2] Knowledge Base System is effective for parsing of GP sentence. Large-scale discriminative two-phase learning algorithms, which ensure the reliable estimation and prevent overfitting, can be used to learn parameters in models with extremely sparse features.[3]Lexical information and knowledge representation can improve the efficiency of parsing. The effective device for imposing structure on lexical information is that of inheritance, both at the object (lexical items) and meta (lexical concepts) levels of lexicon. Lexical semantics theory can utilize a knowledge representation framework to offer a richer, more expressive vocabulary for lexical information, which obviously brings the advancement of knowledge base system.[4]Writing strategy training can be meaningfully provided by artificial intelligence tutoring system which is effective in assessing essay quality and guiding feedback to students with the help of processing algorithms,the development of which must consider a broad array of linguistic, rhetorical, and contextual features.[5]

Machine translation systems need the effectively parsing of GP sentences. MT systems against semantic frame based MT evaluation metrics and objective function can benefit from the semantic knowledge and produce more adequate output. Machine translation evaluation systems in the metrics task can be used to measure the similarity of the system output translations and the reference translations on word sequences.

Statistical machine translation (SMT) performance can be affected by the small amounts of parallel data with the result of both low accuracy (incorrect translations and feature scores) and low coverage (high out-of-vocabulary rates). The bilingual lexicon induction techniques are helpful for learning new translations from comparable corpora, thus improving the coverage. The model’s feature space with translation scores can be estimated over comparable corpora, which brings the improvement of accuracy. [6]Extant Statistical Machine Translation systems embed multiple layers of heuristics and encompass very large numbers of numerical parameters. The phrase-based translation model can be used to decrease the difficulty of analyzing output translations and the various causes of errors.[7]

Based on natural language processing systems, computational linguistics skills and cognitive analysis, related system can be established to efficiently parse GP sentences.

This discussion comprises two parts. The first part will discuss the Shon and Moon’s system for machine learning which is helpful for GP sentence parsing. The second part will parse the GP sentence from multi-perspectives. Context-free grammar, recursive transition network, CYK algorithm, and statistical analysis will be introduced to analyze the processing breakdown of GP sentence.

II. THE SYSTEM FOR MACHINE LEARNING

The peculiarities of linguistic cognition are necessary to be analyzed in the machine learning system. ML should have the ability to process the linguistic data originating from natural languages. For example,the evidence shows that it is hard for children to recover from misinterpretations of temporarily ambiguous phrases. They are reluctant to use late-arriving syntactic evidence to override earlier verb-based cues to structure, and late-developing cognitive control abilities mediate the recovery from GP sentences.[8]World-knowledge about actions designated by verbs and syntactic proficiency is reflected in on-line processing of sentence structure.[9]

Automated knowledge acquisition is an essential aspect involved in machine learning. For example, the automated induction of models and the extraction of interesting patterns from empirical data are concerned in the ML domain. Fuzzy set theory is proved to be helpful for machine learning, data mining, and related fields.[10]

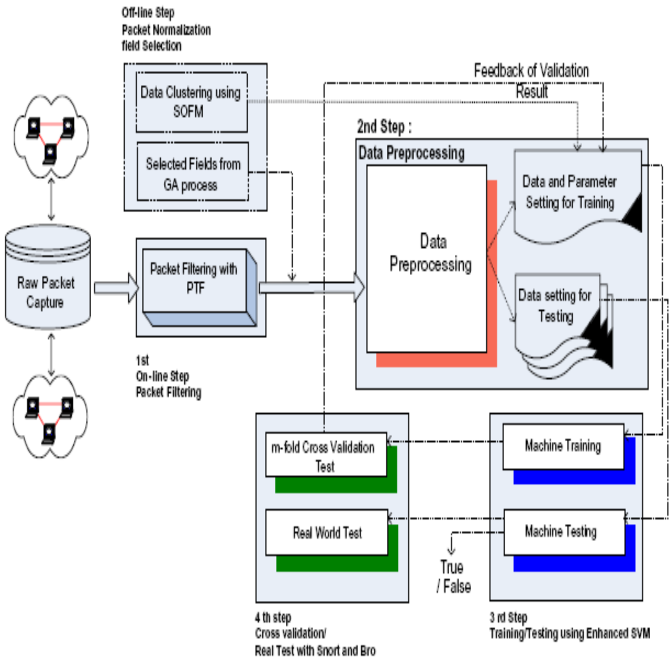


Fig. 1. A Model of Machine Learning Framework

In Fig.1 created by Shon and Moon [11], we can see the system comprises many steps, i.e. on-line processing, off-line processing, feedback of validation, the Enhanced SVM machine learning approaches, cross validation.

On-line processing is the first step, which focuses on a real-time traffic filtering using PTF after the parsing of raw packet capture. PTF is used to drop malformed packets, which can provide efficient work for packet preprocessing and reduce the number of potential attacks of the raw traffic. Off-line processing is involved in data clustering using SOFM and a packet field selection using GA. Both procedures work comfortably before the overall framework. GA chooses the appropriate fields and the filtered packets in real-time select the related fields by the natural evolutionary process. SOFM-based packet clustering can make a lot of packet profiles for SVM learning which can bring more appropriate training data.

In the feedback of validation, the filtered-packets are used as the preprocessing of high detection performance. The relationships between the packets are considered to charge SVM inputs with temporal characteristics. The SVM comprises soft margin SVM (supervised method) and one-class SVM (unsupervised method). The Enhanced SVM machine learning approach, including machine training and testing, inherits the high performance of soft margin SVM and the novelty detection capability of one-class SVM. Cross validation and real test with Snort and Bro is the final step which entails an m-fold cross validation test and real world test.

Shon and Moon's model provides us a general idea about machine learning. An effective system has to be a flexible and adaptable one. That means any effective system should be the product of theoretical and practical application. The improvement of theoretical analysis can improve the system effectively. Therefore, some good algorithms and parsing skills should be involved in improving machine learning system.

From the perspective of theoretical analysis, machine learning depends on the integration of computational technologies and linguistic background. The harmonious development of linguistics and computer science can bring the advancement of machine learning. The parsing of complex sentences, i.e. garden path sentences, will be clearly shown below to present the procedures of machine learning and linguistic cognition. The integration of context-free grammar, recursive transition network, well-formed substring table and CYK algorithm can be used as an effective method of computational linguistics for application.

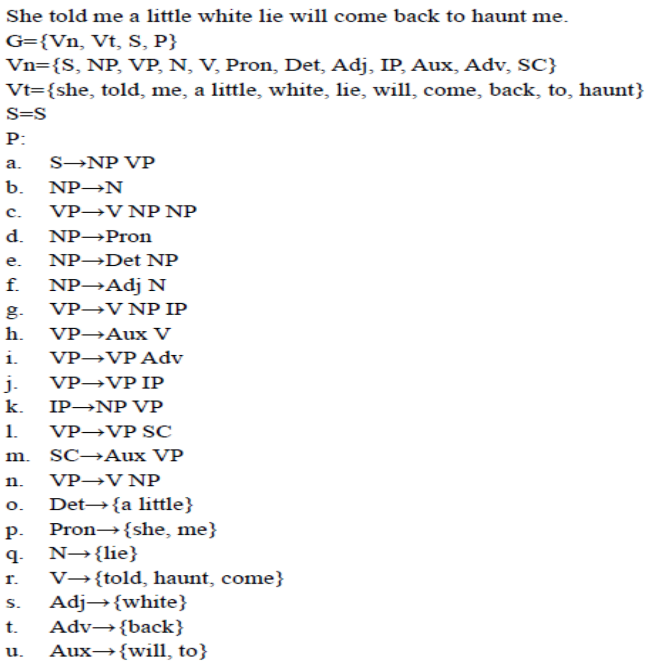
III. MULTI-PARSING OF GP SENTENCE

The processing of GP sentences needs the effective involvement of formal language, e.g. context free grammar, and the algorithms, e.g. recursive transition network and CYK algorithm.

A. *CFG-Based Processing of GP Sentence*

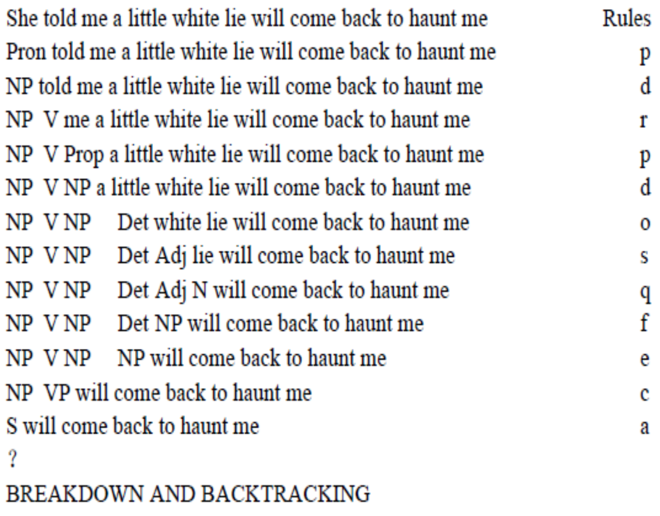
Context free grammar generally comprises nonterminal symbol (V) and terminal symbol (w), and all the production rules belong to the structure of "V→w". Sometimes, "w" can refer to the nonterminal strings or empty strings. The situation in which the production rules are available regardless of the context of a nonterminal is called context free. The single "V" can be replaced by "w" according to the production rules until nearly all the rules are used for the processing. If the ultimate nonterminal symbol "S" can be replaced by the terminal symbols on the right and all the strings have been parsed successfully, the processing is acceptable. If part of the strings are left by system without being parsed, processing has to return to the crossroad to choose the alternative which can lead to the full parsing of all the strings. The backtracking obviously brings the processing breakdown which is the particular effect of GP sentence.

Example 1: She told me a little white lie will come back to haunt me.



She told me a little white lie will come back to haunt me.
G={Vn, Vt, S, P}
Vn={S, NP, VP, N, V, Pron, Det, Adj, IP, Aux, Adv, SC}
Vt={she, told, me, a little, white, lie, will, come, back, to, haunt}
S=S
P:
a. S→NP VP
b. NP→N
c. VP→V NP NP
d. NP→Pron
e. NP→Det NP
f. NP→Adj N
g. VP→V NP IP
h. VP→Aux V
i. VP→VP Adv
j. VP→VP IP
k. IP→NP VP
l. VP→VP SC
m. SC→Aux VP
n. VP→V NP
o. Det→{a little}
p. Pron→{she, me}
q. N→{lie}
r. V→{told, haunt, come}
s. Adj→{white}
t. Adv→{back}
u. Aux→{will, to}

According to the CFG above, we can find the grammar is defined by the four-parameters "Vn, Vt, S, P"and sub-language defined by "G". "Vn" means a finite set of non-terminal symbols; "Vt", a finite set of terminal symbols; "S", start symbol; "P", productions of grammar which shows the relationships between non-terminal and terminal symbols. All the rules are available during the processing. Please see the parsing by means of the grammar.



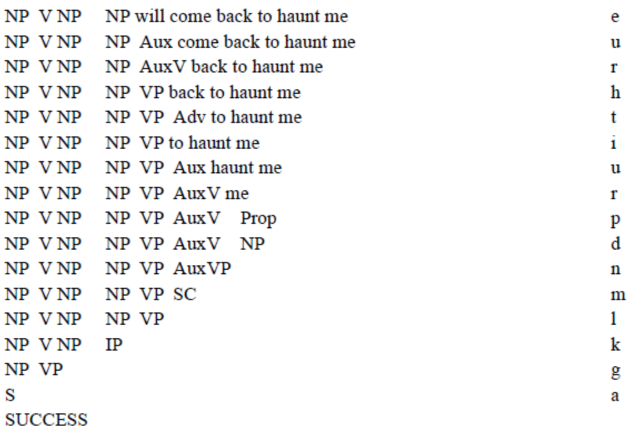
She told me a little white lie will come back to haunt me	Rules
Pron told me a little white lie will come back to haunt me	p
NP told me a little white lie will come back to haunt me	d
NP V me a little white lie will come back to haunt me	r
NP V Prop a little white lie will come back to haunt me	p
NP V NP a little white lie will come back to haunt me	d
NP V NP Det white lie will come back to haunt me	o
NP V NP Det Adj lie will come back to haunt me	s
NP V NP Det Adj N will come back to haunt me	q
NP V NP Det NP will come back to haunt me	f
NP V NP NP will come back to haunt me	e
NP VP will come back to haunt me	c
S will come back to haunt me	a
?	
BREAKDOWN AND BACKTRACKING	

TABLE I. THE STATISTICAL ANALYSIS OF "TELL+ME" TYPE

Type	Observed	Expected	Deviation	D2	D2/E
IP	24	136	-112	12544	92.24
NP	148	136	12	144	1.06
Total	272	272			93.30

In the BNCweb database (http://bncweb.lancs.ac.uk/), we can obtain the statistical information about "tell+me+any article" type in which both "[(that) any article+NP]IP" and "[any article+NP]NP" are acceptable. The corpus shows that "Your query 'tell me' returned 5284 hits in 1213 different texts (98,313,429 words [4,048 texts]; frequency: 53.75 instances per million words), sorted on position +1 with tag-restriction any article (272 hits)". According to the corpus, the statistical number of "[(that) any article+NP]IP" type is 24, and the other NP type is 148.

According to the nonparametric statistics, if df=1 and X^2=93.30, then $p<.05$, which means the significant differences between the types. Since the type of NP has much higher observed frequency than the other, NP type is considered to be the prototype by means of which system automatically parses the strings involved. If the prototypical NP type is rejected by system during the processing, the optional IP type will be started, resulting in the GP effect. Please see the successful processing in which IP type is accepted successfully.



NP V NP	NP will come back to haunt me	e
NP V NP	NP Aux come back to haunt me	u
NP V NP	NP AuxV back to haunt me	r
NP V NP	NP VP back to haunt me	h
NP V NP	NP VP Adv to haunt me	t
NP V NP	NP VP to haunt me	i
NP V NP	NP VP Aux haunt me	u
NP V NP	NP VP AuxV me	r
NP V NP	NP VP AuxV Prop	p
NP V NP	NP VP AuxV NP	d
NP V NP	NP VP AuxVP	n
NP V NP	NP VP SC	m
NP V NP	NP VP	l
NP V NP	IP	k
NP VP		g
S		a
SUCCESS		

In the diagram, we can find the fact that "a little white lie" is considered NP used as the subject of IP, by which the GP sentence is successfully parsed. Besides the CFG and the tree diagram, Recursive Transition Network (RTN) is helpful for the parsing and processing of GP sentence.

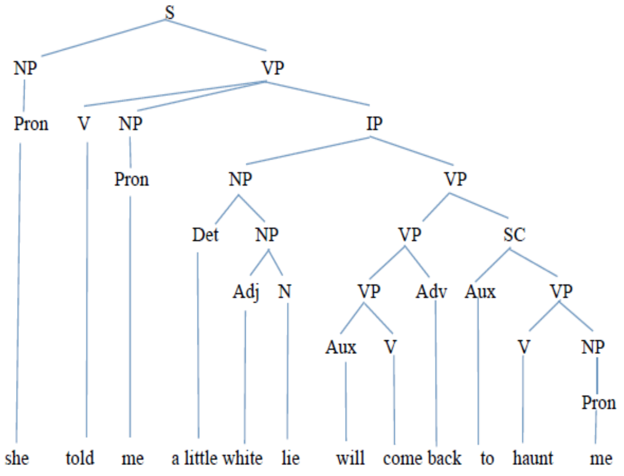


Fig. 2. Tree diagram of example 1

B. RTN-Based Processing of GP Sentence

An RTN usually comprises one S net and some subnets created for the recursive transition of different sub-strings. In

.

Example 1, the local ambiguity potential of structure asks the network effective to parse the sentence. That means both NP subnet and VP subnet should have convincing description of what function "a little white lie" should be, or can be. Please see RTN of Example 1.

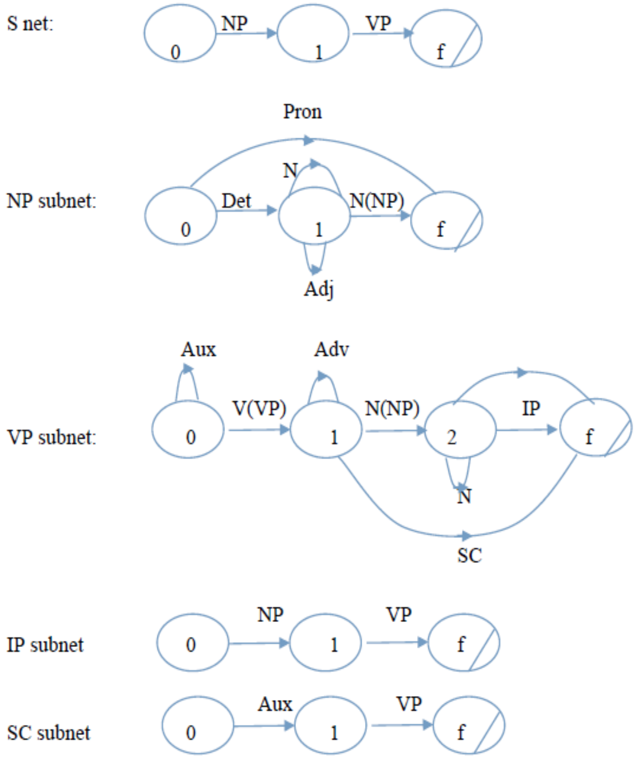


Fig. 3. The RTN of example 1

In Fig. 3, the NP subnet and VP subnet are complexer than the other nets. According to the Table 1, the structure of "She told me a little white lie" is parsed as follows.

```
(ROOT
  (S
    (NP (PRP She))
    (VP (VBD told)
      (S
        (NP (PRP me))
        (NP (DT a) (JJ little) (JJ white) (NN lie)))))
```

The processing above correspondingly brings the backtracking. Please see the details of breakdown processed by means of RTN algorithm.

In the processing, we can see that not all sub-strings are parsed successfully, which means the failure of processing. According to Figure 3, both the structures of "[[she]NP [[told]VP [me]NP [a little white lie]NP]VP]S..." and "[[she]NP [[told]V [me]NP [a little white lie ...]IP]VP]S" are partially acceptable during the processing, thus resulting in the local ambiguity at the point of "a little white lie".

```
She told me a little white lie will come back to haunt me
<S/0, She told me a little white lie will come back to haunt me, >
<NP/0, She told me a little white lie will come back to haunt me, S/1: >
<NP/f, told me a little white lie will come back to haunt me, S/1: >
<VP/0, told me a little white lie will come back to haunt me, S/f: >
<VP/1, me a little white lie will come back to haunt me, S/f: >
<NP/0, me a little white lie will come back to haunt me, VP/1: S/f: >
<NP/1, a little white lie will come back to haunt me, VP/1: S/f: >
<NP/f, a little white lie will come back to haunt me, VP/1: S/f: >
<NP/0, a little white lie will come back to haunt me, VP/2: S/f: >
<NP/1, white lie will come back to haunt me, VP/2: S/f: >
<NP/1, lie will come back to haunt me, VP/2: S/f: >
<NP/f, will come back to haunt me, VP/2: S/f: >
<VP/f, will come back to haunt me, S/f: >
<S/f, will come back to haunt me, >
<, will come back to haunt me, >
?
BREAKDOWN AND BACKTRACKING
```

Since the first structure with high statistical frequency is proved to be illegal above, the alternative structure is started subsequently. The shift obviously brings the cognitive overburden and effect of garden path. The effective processing of unpreferred structure is shown as follows.

```
<NP/f, a little white lie will come back to haunt me, VP/1: S/f: >
<NP/f, a little white lie will come back to haunt me, VP/1: S/f: >
<IP/0, a little white lie will come back to haunt me, VP/2: S/f: >
<IP/1, a little white lie will come back to haunt me, VP/2: S/f: >
<NP/0, a little white lie will come back to haunt me, IP/1: VP/2: S/f: >
<NP/1, white lie will come back to haunt me, IP/1: VP/2: S/f: >
<NP/1, lie will come back to haunt me, IP/1: VP/2: S/f: >
<NP/f, will come back to haunt me, IP/1: VP/2: S/f: >
<VP/0, will come back to haunt me, IP/f: VP/2: S/f: >
<VP/1, come back to haunt me, IP/f: VP/2: S/f: >
<VP/1, back to haunt me, IP/f: VP/2: S/f: >
<VP/2, to haunt me, IP/f: VP/2: S/f: >
<SC/0, to haunt me, VP/f: IP/f: VP/2: S/f: >
<SC/1, haunt me, VP/f: IP/f: VP/2: S/f: >
<SC/1, haunt me, VP/f: IP/f: VP/2: S/f: >
<VP/0, haunt me, SC/f: VP/f: IP/f: VP/2: S/f: >
<VP/1, me, SC/f: VP/f: IP/f: VP/2: S/f: >
<NP/0, me, VP/2: SC/f: VP/f: IP/f: VP/2: S/f: >
<NP/f,    , VP/2: SC/f: VP/f: IP/f: VP/2: S/f: >
<VP/f,    ,SC/f: VP/f: IP/f: VP/2: S/f: >
<SC/f, ,VP/f: IP/f: VP/2: S/f: >
<VP/f, ,IP/f: VP/2: S/f: >
<IP/f, ,VP/2: S/f: >
<VP/f, , S/f: >
<S/f, , >
<, , >
SUCCESS
```

In the processing above, all the sub-strings are included and parsed comprehensively.The procedures reflect the fact that the GP sentence is grammatical correct despite of processing breakdown and backtracking. More details of the decoding algorithm can be analyzed in CYK.

C. CYK-Based Processing of GP Sentence

CYK algorithm is effective for the processing of GP sentence. It can clearly show the procedures of backtracking and present the possibility that computational linguistic knowledge functions as an efficient method providing the helpful hints to alleviate the effect of garden path.

In Table 2, we can see all the procedure of processing breakdown. At the point (7, 6), the CFG rule of "NP-->Adj N" is started and the sub-string "white lie" is parsed. At the point (7, 3), another CFG rule of "NP-->Det NP" is accepted and the sub-string "a little white lie" is rewritten into another NP. At the point(7, 1), the rules of "VP-->V NP NP" and "NP-->Pron" are available. The sub-string of "told me a little white lie" is parsed as VP. At the point(7, 0), the rules of "NP-->Pron" and "S-->NP VP" are provided. The sub-strings of "she told me a little white lie" are processed clearly. The original processing result is obtained. With the appearance of other sub-strings of "will come back to haunt me" which needs a NP as a SUBJECT, system has to re-analyze the original result. Thus, GP effect comes into being. Please see the non-well-formed sub-string table which is created on the basis of Table 2.

TABLE II. THE BREAKDOWN MATRIX OF EXAMPLE 1

	1	2	3	4	5	6	7	8	9	10	11	12	13
0	{P}	{}	{}	{}	{}	{}	{S}	{}	{}	{}	{}	{}	{?}
1		{V}	{}	{}	{}	{}	{VP}	{}	{}	{}	{}	{}	{}
2			{P}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{}
3				{D}	{D}	{}	{NP}	{}	{}	{}	{}	{}	{}
4					{A}	{}	{}	{}	{}	{}	{}	{}	{}
5						{A}	{NP}	{}	{}	{}	{}	{}	{}
6							{N}	{}	{}	{}	{}	{}	{}
7								{A}	{VP}	{VP}	{}	{}	{VP}
8									{V}	{}	{}	{}	{}
9										{A}	{}	{}	{}
10											{A}	{}	{SC}
11												{V}	{VP}
12													{P}

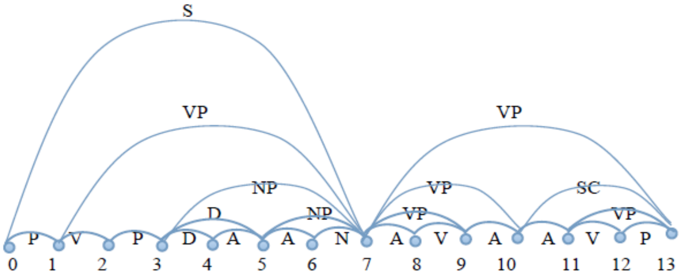


Fig. 4. Non-well-formed sub-string table of breakdown

In Fig. 4, the turning point is 7 in which the processing is cut into S+VP. The result is as same as the situation in Table 2 in which S is obtained at the point(7, 0) and VP is got at the point(13, 7). Since no effective CFG rule used to rewrite S+VP, the sub-string table is not well-formed, meaning the failure of processing. If system can backtrack to the turning point and apply the rule "VP-->V NP IP" rather than "VP-->V NP NP", the whole processing procedures can be attractively presented in Table 3.

TABLE III. THE PROCESSING MATRIX OF EXAMPLE 1

	1	2	3	4	5	6	7	8	9	10	11	12	13
0	{P}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{S}
1		{V}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{VP}
2			{P}	{}	{}	{}	{}	{}	{}	{}	{}	{}	{}
3				{D}	{D}	{}	{NP}	{}	{}	{}	{}	{}	{IP}
4					{A}	{}	{}	{}	{}	{}	{}	{}	{}
5						{A}	{NP}	{}	{}	{}	{}	{}	{}
6							{N}	{}	{}	{}	{}	{}	{}
7								{A}	{VP}	{VP}	{}	{}	{VP}
8									{V}	{}	{}	{}	{}
9										{A}	{}	{}	{}
10											{A}	{}	{SC}
11												{V}	{VP}
													{P}

In Table 3, the CFG rules, e.g. "S-->P(NP) VP", "VP-->V P(N) IP", "IP-->NP VP", "VP-->VP SC" etc. are available and the sentence can be effectively parsed. The well-formed sub-string table created from Table 3 can be shown in the Figure 5. The procedures of parsing reflect the fact that all the strings are successfully processed, and correspondingly, the algorithm used for the system to improve the efficiency can be presented as follows.

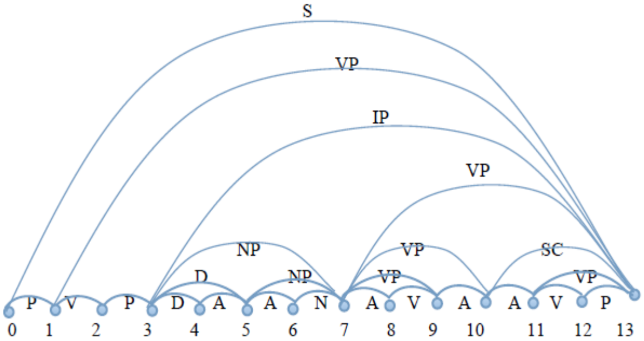
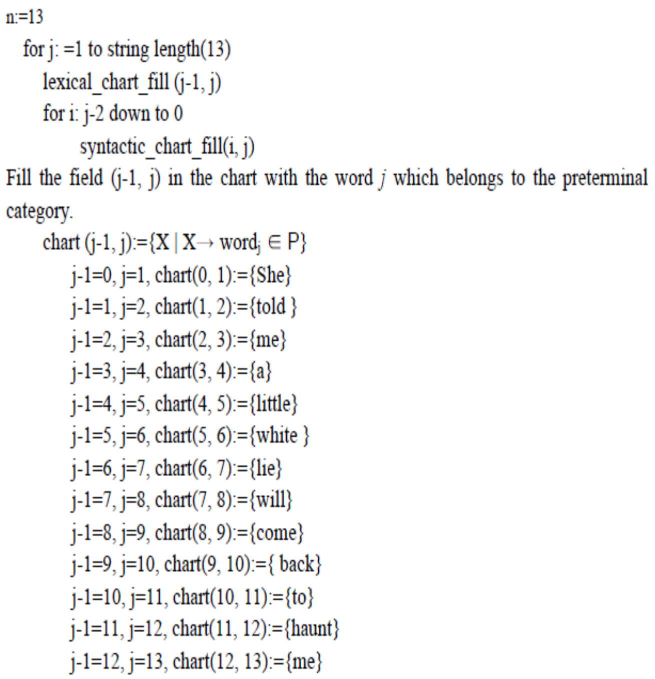


Fig. 5. Well-formed sub-string table of parsing

In the algorithm, two parameters (i & j) are provided to present the procedures. The parameter j is used to construct the chart structure and the parameter i is used to built the grammatical structure.

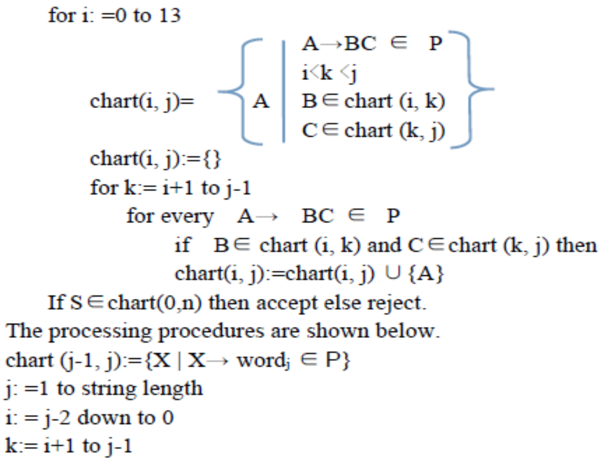
The first step is to construct the chart structure. The sentence has the string length of 13, i.e. 13 words are involved in the processing. The duration of j is 1-13. Then the first word “she”can be drawn by the chart (0, 1); the second word “told”, by the chart(1,2); the third word “me”, by the chart(2,3);...the last word, by the chart(12, 13).



```
n:=13
  for j: =1 to string length(13)
    lexical_chart_fill (j-1, j)
    for i: j-2 down to 0
        syntactic_chart_fill(i, j)
Fill the field (j-1, j) in the chart with the word j which belongs to the preterminal category.
    chart (j-1, j):={X | X→ word_j ∈ P}
        j-1=0, j=1, chart(0, 1):={She}
        j-1=1, j=2, chart(1, 2):={told }
        j-1=2, j=3, chart(2, 3):={me}
        j-1=3, j=4, chart(3, 4):={a}
        j-1=4, j=5, chart(4, 5):={little}
        j-1=5, j=6, chart(5, 6):={white }
        j-1=6, j=7, chart(6, 7):={lie}
        j-1=7, j=8, chart(7, 8):={will}
        j-1=8, j=9, chart(8, 9):={come}
        j-1=9, j=10, chart(9, 10):={ back}
        j-1=10, j=11, chart(10, 11):={to}
        j-1=11, j=12, chart(11, 12):={haunt}
        j-1=12, j=13, chart(12, 13):={me}
```

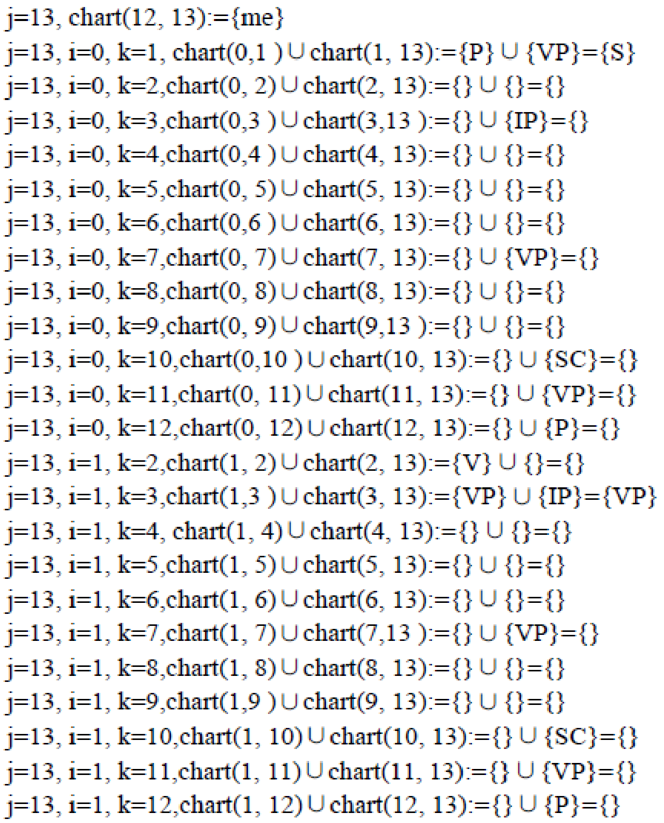
The second step is the construction of syntactic grammar. Since at least two basic charts can constitute one basic syntactic sector, the value duration of i is j-2≧0. Thus, the created syntactic rules of symbols cover the string from i to j, and the algorithm with i and j can be obtained. The basic construction is “syntactic_chart_fill(i, j)”. Please see the algorithm presentation.

In the algorithm, the parameter k input between i and j is introduced to show the dynamic forms of analysis (k:=i+1 to j-1). With the help of i, j, and k, the processing algorithm can be smoothly run, and all the processing procedures based on well-formed sub-string table can be expressed clearly.

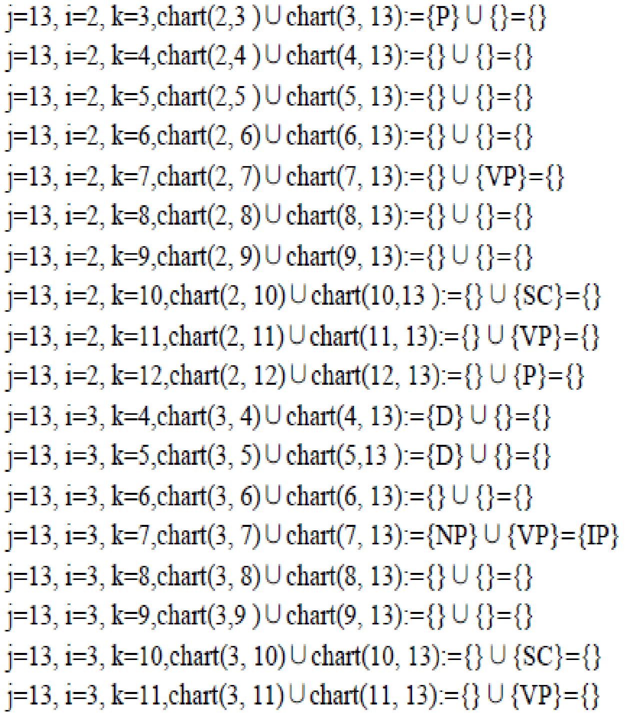


```
    for i: =0 to 13
        $\mathrm{chart}(i, j)= \left\{ A \,\middle|\, \begin{array}{l} A\to BC \ \in \ P \\ i<k<j \\ B\in \mathrm{chart}\ (i, k) \\ C\in \mathrm{chart}\ (k, j) \end{array} \right\}$
        chart(i, j):={}
        for k:= i+1 to j-1
            for every  A→  BC ∈  P
                if  B∈ chart (i, k) and C∈chart (k, j) then
                chart(i, j):=chart(i, j) ∪ {A}
    If S∈chart(0,n) then accept else reject.
The processing procedures are shown below.
chart (j-1, j):={X | X→ word_j ∈ P}
j: =1 to string length
i: = j-2 down to 0
k:= i+1 to j-1
```

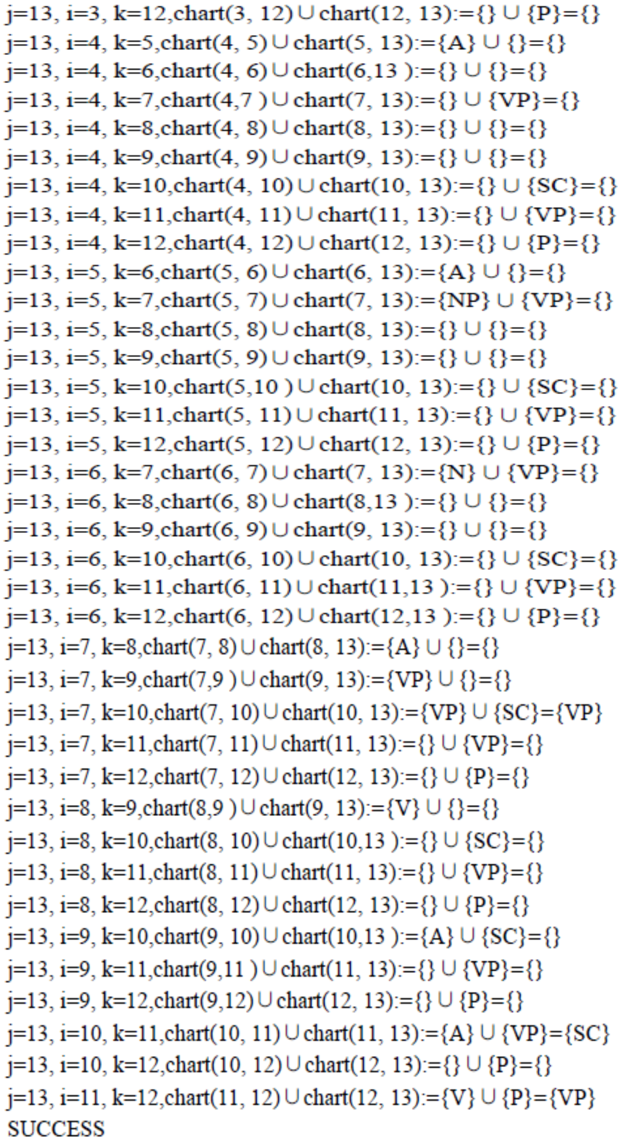
Since the procedures are complex and long, we choose one of the processing procedures (j=13, 11≧i≧0, 12≧k≧1) to show how the algorithm works to parse the sentence.



```
j=13, chart(12, 13):={me}
j=13, i=0, k=1, chart(0,1 )∪chart(1, 13):={P} ∪ {VP}={S}
j=13, i=0, k=2,chart(0, 2)∪chart(2, 13):={} ∪ {}={}
j=13, i=0, k=3,chart(0,3 )∪chart(3,13 ):={} ∪ {IP}={}
j=13, i=0, k=4,chart(0,4 )∪chart(4, 13):={} ∪ {}={}
j=13, i=0, k=5,chart(0, 5)∪chart(5, 13):={} ∪ {}={}
j=13, i=0, k=6,chart(0,6 )∪chart(6, 13):={} ∪ {}={}
j=13, i=0, k=7,chart(0, 7)∪chart(7, 13):={} ∪ {VP}={}
j=13, i=0, k=8,chart(0, 8)∪chart(8, 13):={} ∪ {}={}
j=13, i=0, k=9,chart(0, 9)∪chart(9,13 ):={} ∪ {}={}
j=13, i=0, k=10,chart(0,10 )∪chart(10, 13):={} ∪ {SC}={}
j=13, i=0, k=11,chart(0, 11)∪chart(11, 13):={} ∪ {VP}={}
j=13, i=0, k=12,chart(0, 12)∪chart(12, 13):={} ∪ {P}={}
j=13, i=1, k=2,chart(1, 2)∪chart(2, 13):={V} ∪ {}={}
j=13, i=1, k=3,chart(1,3 )∪chart(3, 13):={VP} ∪ {IP}={VP}
j=13, i=1, k=4, chart(1, 4)∪chart(4, 13):={} ∪ {}={}
j=13, i=1, k=5,chart(1, 5)∪chart(5, 13):={} ∪ {}={}
j=13, i=1, k=6,chart(1, 6)∪chart(6, 13):={} ∪ {}={}
j=13, i=1, k=7,chart(1, 7)∪chart(7,13 ):={} ∪ {VP}={}
j=13, i=1, k=8,chart(1, 8)∪chart(8, 13):={} ∪ {}={}
j=13, i=1, k=9,chart(1,9 )∪chart(9, 13):={} ∪ {}={}
j=13, i=1, k=10,chart(1, 10)∪chart(10, 13):={} ∪ {SC}={}
j=13, i=1, k=11,chart(1, 11)∪chart(11, 13):={} ∪ {VP}={}
j=13, i=1, k=12,chart(1, 12)∪chart(12, 13):={} ∪ {P}={}
```



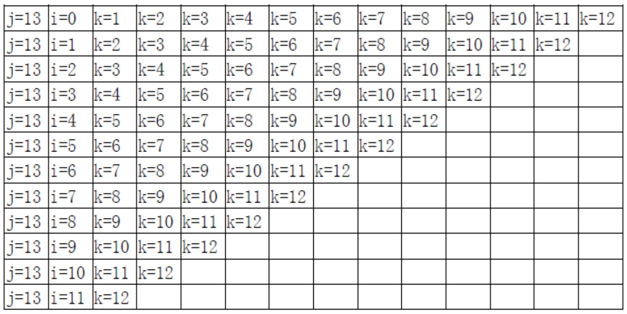
```
j=13, i=2, k=3,chart(2,3 )∪chart(3, 13):={P} ∪ {}={}
j=13, i=2, k=4,chart(2,4 )∪chart(4, 13):={} ∪ {}={}
j=13, i=2, k=5,chart(2,5 )∪chart(5, 13):={} ∪ {}={}
j=13, i=2, k=6,chart(2, 6)∪chart(6, 13):={} ∪ {}={}
j=13, i=2, k=7,chart(2, 7)∪chart(7, 13):={} ∪ {VP}={}
j=13, i=2, k=8,chart(2, 8)∪chart(8, 13):={} ∪ {}={}
j=13, i=2, k=9,chart(2, 9)∪chart(9, 13):={} ∪ {}={}
j=13, i=2, k=10,chart(2, 10)∪chart(10,13 ):={} ∪ {SC}={}
j=13, i=2, k=11,chart(2, 11)∪chart(11, 13):={} ∪ {VP}={}
j=13, i=2, k=12,chart(2, 12)∪chart(12, 13):={} ∪ {P}={}
j=13, i=3, k=4,chart(3, 4)∪chart(4, 13):={D} ∪ {}={}
j=13, i=3, k=5,chart(3, 5)∪chart(5,13 ):={D} ∪ {}={}
j=13, i=3, k=6,chart(3, 6)∪chart(6, 13):={} ∪ {}={}
j=13, i=3, k=7,chart(3, 7)∪chart(7, 13):={NP} ∪ {VP}={IP}
j=13, i=3, k=8,chart(3, 8)∪chart(8, 13):={} ∪ {}={}
j=13, i=3, k=9,chart(3,9 )∪chart(9, 13):={} ∪ {}={}
j=13, i=3, k=10,chart(3, 10)∪chart(10, 13):={} ∪ {SC}={}
j=13, i=3, k=11,chart(3, 11)∪chart(11, 13):={} ∪ {VP}={}
```



```
j=13, i=3, k=12,chart(3, 12)∪chart(12, 13):={}∪{P}={}
j=13, i=4, k=5,chart(4, 5)∪chart(5, 13):={A}∪{}={}
j=13, i=4, k=6,chart(4, 6)∪chart(6,13 ):={}∪{}={}
j=13, i=4, k=7,chart(4,7 )∪chart(7, 13):={}∪{VP}={}
j=13, i=4, k=8,chart(4, 8)∪chart(8, 13):={}∪{}={}
j=13, i=4, k=9,chart(4, 9)∪chart(9, 13):={}∪{}={}
j=13, i=4, k=10,chart(4, 10)∪chart(10, 13):={}∪{SC}={}
j=13, i=4, k=11,chart(4, 11)∪chart(11, 13):={}∪{VP}={}
j=13, i=4, k=12,chart(4, 12)∪chart(12, 13):={}∪{P}={}
j=13, i=5, k=6,chart(5, 6)∪chart(6, 13):={A}∪{}={}
j=13, i=5, k=7,chart(5, 7)∪chart(7, 13):={NP}∪{VP}={}
j=13, i=5, k=8,chart(5, 8)∪chart(8, 13):={}∪{}={}
j=13, i=5, k=9,chart(5, 9)∪chart(9, 13):={}∪{}={}
j=13, i=5, k=10,chart(5,10 )∪chart(10, 13):={}∪{SC}={}
j=13, i=5, k=11,chart(5, 11)∪chart(11, 13):={}∪{VP}={}
j=13, i=5, k=12,chart(5, 12)∪chart(12, 13):={}∪{P}={}
j=13, i=6, k=7,chart(6, 7)∪chart(7, 13):={N}∪{VP}={}
j=13, i=6, k=8,chart(6, 8)∪chart(8,13 ):={}∪{}={}
j=13, i=6, k=9,chart(6, 9)∪chart(9, 13):={}∪{}={}
j=13, i=6, k=10,chart(6, 10)∪chart(10, 13):={}∪{SC}={}
j=13, i=6, k=11,chart(6, 11)∪chart(11,13 ):={}∪{VP}={}
j=13, i=6, k=12,chart(6, 12)∪chart(12,13 ):={}∪{P}={}
j=13, i=7, k=8,chart(7, 8)∪chart(8, 13):={A}∪{}={}
j=13, i=7, k=9,chart(7,9 )∪chart(9, 13):={VP}∪{}={}
j=13, i=7, k=10,chart(7, 10)∪chart(10, 13):={VP}∪{SC}={VP}
j=13, i=7, k=11,chart(7, 11)∪chart(11, 13):={}∪{VP}={}
j=13, i=7, k=12,chart(7, 12)∪chart(12, 13):={}∪{P}={}
j=13, i=8, k=9,chart(8,9 )∪chart(9, 13):={V}∪{}={}
j=13, i=8, k=10,chart(8, 10)∪chart(10,13 ):={}∪{SC}={}
j=13, i=8, k=11,chart(8, 11)∪chart(11, 13):={}∪{VP}={}
j=13, i=8, k=12,chart(8, 12)∪chart(12, 13):={}∪{P}={}
j=13, i=9, k=10,chart(9, 10)∪chart(10,13 ):={A}∪{SC}={}
j=13, i=9, k=11,chart(9,11 )∪chart(11, 13):={}∪{VP}={}
j=13, i=9, k=12,chart(9,12)∪chart(12, 13):={}∪{P}={}
j=13, i=10, k=11,chart(10, 11)∪chart(11, 13):={A}∪{VP}={SC}
j=13, i=10, k=12,chart(10, 12)∪chart(12, 13):={}∪{P}={}
j=13, i=11, k=12,chart(11, 12)∪chart(12, 13):={V}∪{P}={VP}
SUCCESS
```

From the procedures above, we can see the details of the processing which includes 78 steps if j=13. When i=0, k expands from 1 to 12(12 steps); when i=1, k expands from 2 to 12 (11 steps); when i=2, k expands from 3 to 12 (10 steps)...When i=11, k is 12(1steps). All the steps involved can be obtained as follows: 12+11+10+9...+4+3+2+1=78. The calculations of steps are shown in Table 4.

TABLE IV. THE PROCESSING PROCEDURES (J=13)



j=13	i=0	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12
j=13	i=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12	
j=13	i=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12		
j=13	i=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12			
j=13	i=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12				
j=13	i=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12					
j=13	i=6	k=7	k=8	k=9	k=10	k=11	k=12						
j=13	i=7	k=8	k=9	k=10	k=11	k=12							
j=13	i=8	k=9	k=10	k=11	k=12								
j=13	i=9	k=10	k=11	k=12									
j=13	i=10	k=11	k=12										
j=13	i=11	k=12											

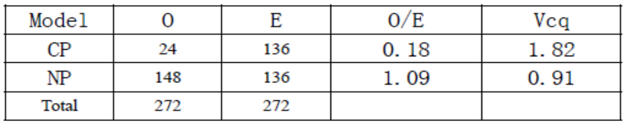
D. The Statistical Analysis of GP Effect

Effect of garden path can be calculated based on the statistical analysis. The formula created to analyze the confusion quotient is as follows:

$$V_{cq}=\frac{\sum_{i=1}^{n}(1-\frac{Oi-Ei}{Ei})}{n}=\frac{1}{n}\sum_{i=1}^{n}(2-\frac{Oi}{Ei})$$

According to the details of parameters, Vcq is used to show the confusion degree and cq means “confusion quotient”. The letter O is the abbreviation of “observed frequencies” during the processing. The letter E is the abbreviation of “expected frequencies”. The letter n is the abbreviation of “number”, meaning the numbers of peculiarities. The letter i means the unit of peculiarity. The complexer the structure of garden path sentence is, the higher the confusion quotient is. The value of 1 is the axis of confusion quotient. The GP effect appears when 2≧Vcq≧1. If the value of Vcq is less than 1, whole ambiguity appears rather than the partial ambiguity which is the result of GP effect.[12]

TABLE V. THE CONFUSION QUOTIENT OF CP/NP AMBIGUITY



Model	O	E	O/E	Vcq
CP	24	136	0. 18	1. 82
NP	148	136	1. 09	0. 91
Total	272	272		

In Table 5, we can see the CP model has the value (1.82) and NP model has the value (0.91), which means the NP structure has less confusion quotient and therefore is considered the prototype. If the preferred NP model is replaced by the unpreferred or higher confusion model, the processing breakdown appears and GP phenomenon takes effect. The statistical analysis of GP effect reflects the fact that the ungrammatical preferred structure is always parsed firstly, which brings the backtracking during the processing. The grammatical unpreferred structure is proved to be the replacement of the initial processing. The chronological start of partial ambiguous structures is the potential reason of processing breakdown. The parsing can be shown by the online Stanford Parser which focuses more on statistics. The parsing is as follows:

```
She/PRP          told/VBD
me/PRP           a/DT
little/JJ        white/JJ
lie/NN           will/MD
come/VB          back/RP
to/TO            haunt/VB
me/PRP           ./.
(ROOT
  (S
    (NP (PRP She))
    (VP (VBD told)
      (NP (PRP me))
      (SBAR
        (S
          (NP (DT a) (JJ little) (JJ white) (NN lie))
          (VP (MD will)
          (VP (VB come)
            (PRT (RP back))
            (S
              (VP (TO to)
                (VP (VB haunt)
                  (NP (PRP me)))))))))
    (. .)))
nsubj(told-2, She-1)
root(ROOT-0, told-2)
dobj(told-2, me-3)
det(lie-7, a-4)
amod(lie-7, little-5)
amod(lie-7, white-6)
nsubj(come-9, lie-7)
aux(come-9, will-8)
ccomp(told-2, come-9)
compound:prt(come-9, back-10)
mark(haunt-12, to-11)
xcomp(come-9, haunt-12)
dobj(haunt-12, me-13)
```

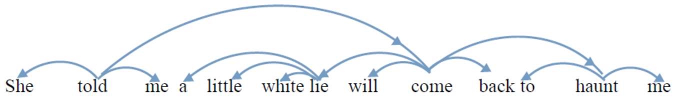


Fig. 6. Universal dependencies of GP sentence from Stanford Parser

From the parsing above, we can learn that Stanford Parser has the potential to deal with some kind of local ambiguity, e.g. tell+CP/NP, based on the Probabilistic Context Free Grammar (PCFG). And the dependency grammar is proved to be useful to semantically analyze the complex sentence. Please see the universal dependencies provided by the parser.

IV. Conclusion

The ambiguous structure of V+Pron+CP/NP has the potential to result in the processing breakdown. If V+Pron+NP structure has high observed frequency and low confusion quotient, the structure is considered the preferred one which is parsed initially and the parsing is ungrammatical. The replacement of ungrammatical preferred structure of V+Pron+NP by the grammatical unpreferred structure of V+Pron+CP can bring the GP effect. The methods of computational linguistics are proved to be effective to explore the phenomenon.

GP effect has been discussed by scholars from different perspectives [13-18], and this local ambiguity phenomenon with processing breakdown and backtracking deserves serious attention, especially for natural language processing. This paper discusses the effect theoretically and draws a conclusion that an effective system needs the help of multiple disciples, including computational linguistics, cognitive linguistics, psychological linguistics, computer science etc. How to balance the needs of rule-based and statistics-based methods is a new challenge for researchers of computational linguistics in future.

Acknowledgment

This research is supported in part by China National Social Science Fund (12FYY019) and China Postdoctoral Fund (2015M570424)

References

[1] M. H. Christiansen and N. Chater, “ Connectionist natural language processing: The state of the art,” Cognitive Science, 1999, 23(10), pp. 417-437.

[2] R. M. Losee, “Natural language processing in support of decision-making: Phrases and part-of-speech tagging,” Information Processing & Management, 2001, 37(11), pp. 769-787.

[3] J. Flanigan, C. Dyer, and J. Carbonell, “Large-Scale Discriminative Training for Statistical Machine Translation Using Held-Out Line Search,” Proceedings of NAACL-HLT, 2013, pp. 248-258.

[4] J. Pustejovsky and B. Boguraev, “Lexical knowledge representation and natural language processing,” Artificial Intelligence, 1993, 63(10), pp. 193-223.

[5] D. S. McNamara, S. A. Crossley, and R. Roscoe, “Natural language processing in an intelligent writing strategy tutoring system,” Behavior Research Methods, 2013, 45(2), pp. 499-515.

[6] A. Irvine and C. Callison-Burch, “Combining bilingual and comparable corpora for low resource machine translation,” Proceedings of the Eighth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2013, August, pp. 262-270.

[7] G. Wisniewski and F. Yvon, “Oracle decoding as a new way to analyze phrase-based machine translation,” Machine Translation, 2013, 27(2), pp. 115-138.

[8] Y. Choi and J. C. Trueswell, “Children’s (in)ability to recover from garden paths in a verb-final language: Evidence for developing control in sentence processing,” Journal of Experimental Child Psychology, 2010, 106(5), pp. 41-61.

[9] E. Malaia, R. B. Wilbur, and C. Weber-Fox, “ERP evidence for telicity effects on syntactic processing in garden-path sentences,” Brain and Language, 2009, 108(3), pp. 145-158.

[10] E. Hüllermeier, “Fuzzy sets in machine learning and data mining,” Applied Soft Computing, 2011, 11(2), pp. 1493-1505.

[11] T. Shon and J. Moon, “A hybrid machine learning approach to network anomaly detection,” Information Sciences, 2007, 177(18), pp. 3799-3821.

[12] J. L. Du, The Asymmetric Information Compensation Hypothesis: Research on confusion Quotient in Garden Path Model, Doctoral Dissertation for Communication University of China, 2013.

[13] J. L, Du and P.F.Yu, “A computational linguistic approach to natural language processing with applications to garden path sentences analysis,” International Journal of Advanced Computer Science & Applications, 2012, 3(9), pp.61-75.

[14] J. L, Du and P.F.Yu, “Predicting Garden Path Sentences Based on Natural Language Understanding System,” International Journal of Advanced Computer Science & Applications, 2012, 3(11), pp.1-7.

[15] J. L, Du, P.F.Yu, and M. L. Li, “Machine Learning from Garden Path Sentences: The Application of computational Linguistics,” International Journal of Emerging Technologies in Learning, 2014, 9(6), pp. 58-62.

[16] Z. G. Shooshtari, and S. Shahri. "Down the Garden Path: An Effective Kind of EFL Grammar Instruction," Procedia Social and Behavioral Sciences, 2014, 98(6), pp.1777–1784.

[17] T. J. Slattery,et al. "Lingering Misinterpretations of Garden Path Sentences Arise from Competing Syntactic Representations," Journal of Memory & Language, 2013, 69(2), pp.104–120.

[18] Y. Choi, "Children’s (In)Ability to recover from garden paths in a verb-final language: evidence for developing control in sentence processing," Journal of Experimental Child Psychology, 2010, 106(1), pp. 41–61.

Application of GLBP Algorithm in the Prediction of Building Energy Consumption

Dinghao Lv[1]
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Bocheng Zhong[2]
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Jing Luo[3]
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

***Abstract*—Using BP neural network in past to predict the energy consumption of the building resulted in some shortcomings. Aiming at these shortages, a new algorithm which combined genetic algorithm with Levenberg-Marquardt algorithm (LM algorithm) was proposed. The proposed algorithm was used to improve the neural network and predict the energy consumption of buildings. First, genetic algorithm was used to optimize the weight and threshold of Artificial Neural Network (ANN). Levenberg-Marquardt algorithm was adopted to optimize the neural network training. Then the predicting model was set up in terms of the main effecting factors of the energy consumption. Furthermore, a public building power consumption data for one month is collected by establishing a monitoring platform to train and test the model. Eventually, the simulation result proved that the proposed model was qualified to predict short-term energy consumption accurately and efficiently.**

***Keywords*—BP Neural network; Building energy consumption; Genetic algorithm; Levenberg-Marquardt algorithm**

I. INTRODUCTION

There are many factors affecting the energy consumption of the building, and it has the characteristics of randomness, time-varying and regionally [1]. It can be divided into three aspects: environment, structure and operation process. For a given area of buildings, the main influence factors of the building energy consumption include regional climate characteristics, buildings, residential environment, building construction and operation management of heating system [2-3]. Because neural network [4] has strong nonlinear, parallel processing ability and robustness, and it does not need to set up complex mathematical model, so it has been favored by the researchers. Now it is widely used in different conditions of architecture energy consumption prediction research.

The traditional algorithm exists the shortcoming of slow convergence speed and easy to fall into local minimum. Bingbing Shi [5] and others find that when using LM algorithm to improve the BP neural network, the BP network training speed is improved obviously, and the prediction error is smaller. Ahua Mu [6] and others by using the method of combining genetic algorithm with BP algorithm (hereinafter referred to as GABP algorithm) makes the prediction error smaller, but the amount of computation is increased and the convergence rate is slow. By the literature [7], under the same training target, the training steps of genetic neural network are 10 times of the LM algorithm, so the training speed of GABP algorithm is not ideal, and it can't satisfy the requirement of energy consumption prediction of real-time online. For this kind of situation, in this article, the GLBP algorithm which combines the GABP algorithm and the LM algorithm is adopted to establish the building energy consumption prediction model.

II. GLBP ALGORITHM

The improved BP neural network is mainly divided into two parts, one part is to optimize the input global variable, and to improve the quality of input variables; the other part is to optimize local variables, and to improve the convergence speed of the algorithm. Therefore, the improved BP neural network (GLBP) has two stages in the calculation: 1) Using the genetic algorithm to determine the approximate optimal approximate solution; 2) The LM algorithm is used to search the approximate solution, until you find the local optimal solution. The specific steps are as follows:

A. Genetic algorithm global optimization [8]

1) According to the network structure, to determine the code length of genetic algorithm. Using real coding, changing the parameter set X and fields into the string structure space S;

2) The fitness function is the only standard to measure the quality of the individual. In order to improve the prediction accuracy, the reciprocal of the error sum of squares of the LM neural network is used as the fitness function. The function is:

$$f(X_i) = \frac{1}{\sum_{i-1}^{k} \left(t_i - y_i\right)^2} \qquad (1)$$

Among them, k is the number of training samples; t is target output; y is the actual output;

3) Determine the genetic strategy, including selected population size n, method of selection, crossover and mutation, and crossover probability P_e and mutation probability P_m;

4) Population initialization, the individual of the population is encoded to calculate the initial fitness valu f(x);

This work was supported by the Shanghai board of education discipline construction under Grant no. XKCZ1212, and also was supported by the Innovation Program of Shanghai University of Engineering Science under Grant No.E1-0903-14-010340

5) Selection of the genetic operation method based on step 3), the genetic operations of selection, crossover and mutation are carried out to form the next generation population;

6) To determine whether the new individual population termination condition, if the condition is satisfied, the search will be stopped, otherwise, return to step 5), until satisfy the termination condition.

B. LM algorithm local optimization [9]

1) Selecting the largest fitness from the group which meet the conditions as the initial weights and thresholds of the training of the LM neural network;

2) Again using the training samples of LM neural network. Input samples, and then calculate, and get a response in the output layer;

3) In accordance with the direction of reducing the actual output and target output error to modify the weights of each neuron. Unlike the BP algorithm, LM algorithm uses two order derivative approximation, and the rate of convergence is much faster than BP algorithm. The LM algorithm allows the error along the direction of increasing development, not easy to fall into local minimum.

4) To determine whether the training results meet the certain indicators, not meet return 3), else end.

III. Building Energy Consumption Prediction Model

Through the analysis of the factors affecting energy consumption of a particular building, considering the human activity is relatively fixed, the equipment operation and the enclosure structure basically remain unchanged all the year, so here mainly consider the influence of the outdoor climate conditions (weather) to energy consumption. According to the characteristics of the particular building energy consumption, selecting temperature, humidity and wind speed as the input layer factors, and selecting the electric energy consumption as the output layer factor. When you want to predict the future energy consumption, you just need the future weather forecast information. Here, the number of input layer neurons is three. The number of output layer neurons is one and the number of the hidden layer also is one. The energy consumption prediction model is built based on this, then the model diagram (three-ten-one) is shown in Fig one.

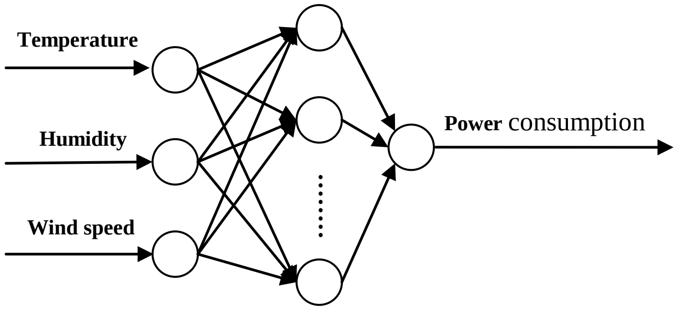


Fig. 1. Energy consumption prediction model

To obtain the experimental data, a university training center is taken as the research object, and then set up the experimental platform. Because of the large proportion of energy consumption, and other forms of energy consumption can be collected by similar methods, the energy consumption is only selected as the collection object.

Energy consumption data acquisition system uses three-tier design: the field monitoring layer, the network communication layer and management layer. The management layer sends out data acquisition command, then transfers the command to each electric meter by network communication layer. Electric meter receives the instruction and checks it, then submits the corresponding energy consumption information to management computer, and finally stores it to the database. The prediction model can obtain the training sample from the database.

IV. Simulation and Analysis

From the database to select three months (ninety days) of raw data. Dividing these sample data into two parts. The first part is as the training sample of neural network and this part includes the data of 93 days. The remaining data is as the test samples. In order to improve the tolerance of the network, add "noise" in each group of training samples, therefore, the number of the training sample is 186.

According to the characteristics of the neural network processing data, the Premnmx function [10] is used to control the training samples normalized between [-1, 1], and the method is as follows:

$$p' = 2\frac{p - p_{\min}}{p_{\max} - p_{\min}} - 1 \qquad (2)$$

Type: p' is the normalized variable; p is the primitive variable; $p_{\max}$ 、 $p_{\min}$ respectively is the maximum and the minimum of the original variables.

Contrast the GABP algorithm with the GLBP algorithm in the simulation experiment, structural parameters of the model are consistent. Parameters of genetic algorithm: population size is $n = 50$; genetic algebra is $gen = 100$; crossover rate is $P_c = 0.5$; the mutation rate is $P_m = 0.01$. Part of neural networks: vector is 0.3; the largest number of iterations is 8000 and the minimum error of the training target is 0.001. The neuron transfer function of the network interface layer adopts s-shaped tangent function, tansig[11], the neuron transfer function of output layer adopts linear function, purelin[12].

Type: The GABP algorithm is an algorithm which combine neural network and genetic algorithm. The GLBP algorithm is proposed in this paper.

MATLAB as the simulation tool, then respectively use the GABP and GLBP algorithm to build energy consumption prediction model. After the simulation experiments, the training error curve of the test results are as follows:

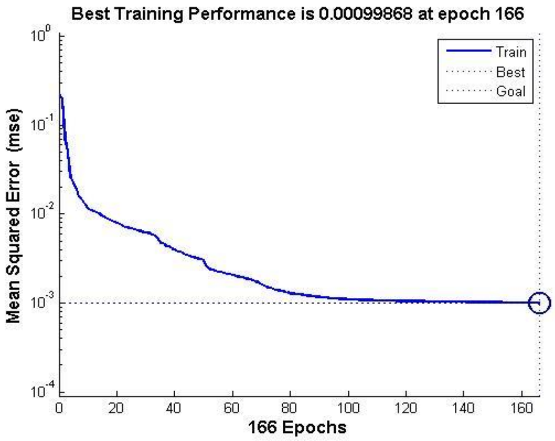


Fig. 2. GLBP training error trend chart

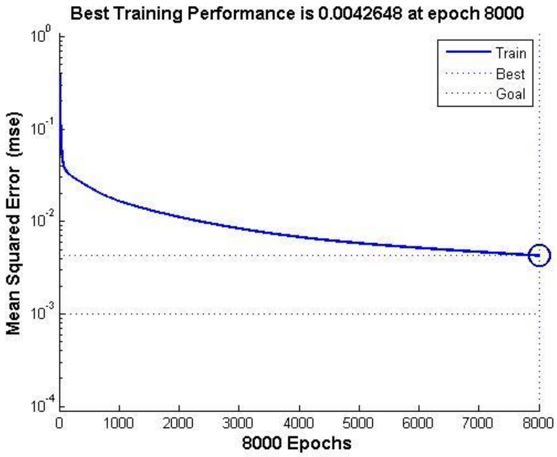


Fig. 3. GABP training error trend chart

The experimental results show that, GLBP algorithm for network training is faster than GABP algorithm, and the number of basic iterative can achieve the training objectives within 200 times, as shown in Fig 2. All of them use the same method which uses genetic algorithm to search the optimal weights and threshold value. However, not only does the GABP algorithm increase the amount of calculation and consume a long time, but also can’t make neural network jump out of the local minimum point, and it can be seen from Fig 3. Even if the number of training reach the maximum (8000 times), network still failed to reach the training target. LM algorithm was used to make up the time of the searching process of genetic algorithm, and increased the speed of network training, finally it can quickly reach the goal of the training. This shows that in this paper the GLBP algorithm can satisfy the requirement of the real-time prediction of energy consumption.

Not only does the Building energy consumption prediction need the speed, but also need to meet a certain precision, to suit the requirements of engineering application. Using the GLBP to predict the energy consumption of the next 7 days, the prediction result is shown in table 1. Overall, the maximum relative error is 3.86%; the average relative error is 1.37%; the prediction accuracy is relatively high, so it could meet the needs of the actual demand the building energy consumption. The forecast result of GLBP is as follows:

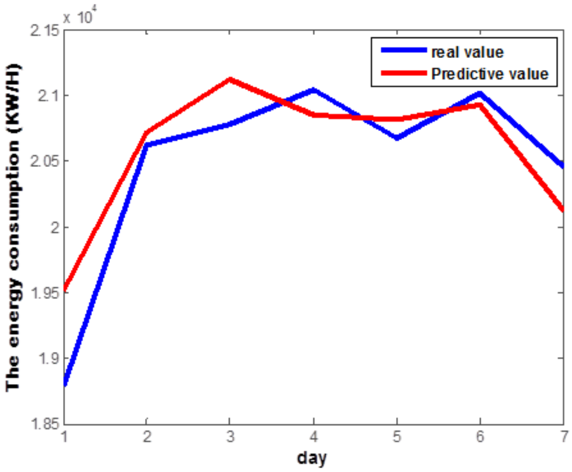


Fig. 4. The forecast results of GLBP

TABLE I. GLBP PREDICTION RESULTS

date	1	2	3	4	5	6	7
real value	18787.0	20616.5	20778.5	21040.5	20673.0	21012.5	20450.5
Predictive value	19515	20716	21117	20850	20815	20923	20119
relative error	3.86%	0.48%	1.63%	0.91%	0.67%	0.42%	1.61%

V. CONCLUSION

According to the requirement of real-time prediction of building energy consumption in short term, this paper proposes genetic algorithm combined with the LM algorithm, namely GLBP algorithm, and this method improves the performance of neural network to predict the building energy consumption. Through the analysis of main influence factors and setting the network parameter, the building energy consumption prediction model is established based on GLBP. In order to verify the feasibility of the model, energy monitoring platform is built to collect the electric energy data, by recording the weather conditions to train and test the model, in order to improve the fault tolerance of the network, join the interference data in the training samples. The results show that the GLBP algorithm training time is short, not easy to fall into local minima, strong generalization ability, and the prediction accuracy can satisfy the requirement of the engineering application.

In the future, not only does it need to predict energy consumption in a short period of time, but also need to predict the energy consumption of a month or even a year. Long-term energy consumption prediction will have a brighter future.

REFERENCES

[1] Zhao, Hai-xiang, and Frédéric Magoulès. "A review on the prediction of building energy consumption." Renewable and Sustainable Energy Reviews 16.6 (2012): 3586-3592.

[2] Mena, R., et al. "A prediction model based on neural networks for the energy consumption of a bioclimatic building." Energy and Buildings 82 (2014): 142-155.

[3] Binshou Tian. Building energy efficiency testing technology (second edition) [M]. Beijing: China building industry press, 2010.8.

[4] Ekici B B, Aksoy U T. Prediction of building energy consumption by using artificial neural networks [J]. Advances in Engineering Software, 2009, 40(5): 356-362.

[5] Bingbing Shi, Zhemin Duan, Zhengjun Lu. Comparative study on the neural network to predict medium and long-term power load [J]. Relay, 2007, 35(23): 43-45.

[6] Ahua Mu, Shaolei Zhou, Zhiqing Liu, etc. By using the genetic algorithm improved BP learning algorithm [J]. The computer simulation, 2005, 22(2): 150-151.

[7] Neto A H, Fiorelli F A S. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption [J]. Energy and Buildings, 2008, 40(12): 2169-2176.

[8] Alsayegh O, Almatar O, Fairouz F, et al. Prediction of the long-term electric power demand under the influence of A/C systems [J]. Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy, 2007, 221(1): 67-75.

[9] Chen Haiying, Guoqiao, Xuli. Hundred meters based on hybrid genetic neural network prediction method [J]. The computer simulation, 2004, 21(2):89-91.

[10] Li Minqiang, Kou Jisong, Lindan, etc. The basic theory of genetic algorithms and application [M]. Beijing: science press, 2002.3.

[11] Zhao Qingyan. Genetic optimization neural network network traffic prediction [J]. Microelectronics and computer, 2013, 30 (3):133-135.

[12] Kůrková V. Kolmogorov's theorem and multilayer neural networks [J]. Neural networks, 1992, 5(3): 501-506.

Smart Transportation Application using Global Positioning System

Nouf Mohammad Al Shammary and Abdul Khader Jilani Saudagar
Information Systems Department, College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

***Abstract*—Significant increase is noticed in the utilization of mobile applications for different purposes in the past decade. These applications can improve any individual's way of life in many aspects such as communication, collaborative work, learning, location services, data collection, exploring, testing and analysis. One of the most interesting mobile applications is using it for tracking by having personal locators. These locators can track children, people on work, the elderly for personal protection etc. The intention behind developing this mobile application is to provide a smart transportation system to it users and to track their movements.**

Some of the essential features of this application are

- **Getting familiar with the shortest path from source to destination in advance.**
- **Aware of approximate time of arrival to destination.**
- **Knowing the capacity of vehicle used for transportation.**
- **Short Message Service.**

Keywords—communication; global positioning system; smart application; tracking; transportation

I. Introduction

Managing transportation is a collaborative activity that is based upon sharing and exchanging information. Transportation system needs an easy and effective way to manage their staff and the registered users who are using the transportation services. Staff needs an effective way to know the shortest path to get to the locations, handle the requests of the users etc. Registered users also need to keep up with the staff movements without waiting too early or causing too much delay. The registered users can inform the staff about their absence using short message service. Providing the above mentioned services collectively and managing them from any location is a challenging problem. Mobile technology can be a solution for the above problem using location based applications which serves managing transportation by navigating maps. In recent years, mobile technology revolution and its applications reached to its climax. This revolution resulted in an aggressive competition between the application developers as the developed mobile applications are similar in concepts and are identical in design but are set apart by unique features, offered services to complement the main idea.

Mobile phones are becoming smaller with modern Graphical User Interfaces (GUIs), and are very powerful. In addition, they are now can be used everywhere and at any time (ubiquitous). The advancements and enhancements the hardware went through have made mobiles smaller in size and more effective and efficient. Furthermore, it allows us to include many types of peripherals that are not limited with a specific number [1]. Nowadays, there are three different methodologies to allow smart mobile phones to be capable of utilizing the location and positioning services: 1) The utilization of Satellite Positioning, 2) The utilization of Wi-Fi Positioning and 3) The utilization of Cellular Positioning. When looking at these positioning methodologies in a general way, it is noticeable that they have different characteristics which can be measured by Accuracy, Precision, Power consumption, Latency and Availability as proposed by the researcher Hightower et. al [2].

Tracking service providers[1] [2] [3] save the information related to locations in their own database and users are capable of accessing this information. It can be accessed and efficiently utilized upon paying over international telecommunication infrastructure. Furthermore, taking advantage of the SMS availability in mobile networks facilitates users to communicate, transfer information from one place to another. In these types of systems, the users will be more limited since this service has limited flexibility with higher cost.

In [4] the work entitled "GIS Based Public Bus Transport Management System for Nairobi City", Otieno and Ngigi introduced a public bus transport management system that utilizes Geographic Information Systems (GIS) and Global Positioning System (GPS) technologies and the system completely depends on these technologies. The system is capable of storing the day-by-day information that a public transport vehicles produce in an intension to help in having an efficient and effective management of public transport. This work also demonstrated the way of how the collected tracking data can be utilized to manage the movements of public transport vehicles when it is added and mixed with other types of data from other sources, for instance, vehicle monitoring inventory.

In [5] the researchers had the aim to achieve the following two objectives:

[1] http://thenextweb.com/location/2010/08/24/geo-positioning-on-the-iphone-4-is-doing-just-fine-without-skyhook-and-google

[2] http://www.engadget.com/2010/09/17/ skyhook-google-forced-motorola-to-drop-our-location-service-de/

- Developing an automatic vehicle location system. In addition, GPS, GSM technologies will be necessary for the purpose of information transmission.
- Sending some of information that is related to the locations of vehicles (the focus was on buses) to children's parents through the utilization of SMS.

Global Navigation Satellite System (GNSS) is the primary system in their project that they are building; it utilizes a satellite that is used for positioning and tracking. GNSS is primarily utilized for monitoring numerous kinds of vehicles, for instance, airplanes, cars (including buses) and other kinds of vehicles. In this paper, they considered buses are the most important and common ways for transportation that allow us to move from one place to another. As a result of that, this is the vehicle that this proposed application will deal with mostly. When passengers coming from different places move or travel from one location to another, there is a chance that they may face different types of problems, for instance, knowing the arriving time of a bus so they can use their time wisely as well as knowing the time when to reach the desired bus stop etc. As described by the researchers who proposed this mobile application, the primary goal of this application is to minimize the time that the parents should wait when they are at the bus stop. This can be achieved through sending some information related to the different locations that the buses pass by during their movements to those parents through the utilization of SMS. The idea of this application is providing a mobile application that will benefit passengers to get to a bus stop without wasting any time. In practice the buses are often late due to several reasons, for instance, intense traffic. It would be a very useful idea to be able to build a web application for observing buses movements in real time. The process of observing buses is performed through utilizing manual methods to get the position of a bus, for instance, signals are manually operated by staff that perform the instructions using telephone from the closest bus stop. So, information related to buses movements that is propagated to the public is dependent on manual process. The proposed system only gives the position of a bus but cannot give the exact position of the bus or where it is in a particular moment. To solve this problem, a GNSS based web application is utilized which provides with the exact location of a bus that will be shown on Google Maps as well as the bus speed. Combining GNSS with a web tools such as Google Maps and web browser offers a cheap and useful bus monitoring system that solves several problems.

Other mobile application that is used to find a bus arrival time is called Ride Systems GPS[3]. This system provides passengers with the next bus time, in other words, the time when their next bus will arrive. The most important feature that this mobile application provides is finding stops. Furthermore, it provides the arrival times for a particular route. This application can be used by Android and iPhone operating systems.

In the report [6] entitled "Wireless Global Positioning System Fleet Tracking System at the University at Albany" prepared by New York State Energy Research and Development Authority, an overview of the project was proposed at the University at Albany that is introduced to make interesting ways to interact with transportation facilities which produces a more attractive option to deal with these facilities. The system includes implementing a GPS Tracking System on the University bus station. This system sends the different locations that the bus passes by to the users of the system. This propagation will be through broadcasting this information to passengers via the internet and a smart phone application. In 2009, the university made a survey related to transportation of the students and staff of the university. From the survey analysis, the results were as follows:

- 73% of the staff working there and 39% of the students studying there have used primarily a private vehicle for transportation to and from the university.
- People who have actually utilized the University of Albany bus daily were just 18% of the students and 2% of the staff.

Participants to the proposed survey have mentioned that the biggest obstacle behind using the university buses was related to the "convenience" factor. After utilizing the GPS real-time tracking system and installing it, it has been noticed that there was a small increase in the percentage of passengers that utilized the transportation of University of Albany buses.

Mobile tracking applications are very popular nowadays. One of the most well-known location tracking systems is the save and sound system [7]. This system enables parents to monitor their children's locations. The child's phone continuously streams location information to the parent's phone. A secure zone will be created beyond which a child may not travel. If the child leaves this zone, both child and parent receive an audible alert, and the parent can communicate with the child by voice over the phone.

However, the Federal Communications Commission (FCC) of the United States has mandated all wireless service providers to provide an E-911 service whereby the mobile user location is reported when they dial an emergency call (911). This development has helped many location based services (LBS) [8]. Various application systems have emerged for mobile location systems such as location-based mobile tourist services, location-based game and location-based speaker segmentation [9].

Marco Anisetti and colleagues [10] agreed that GPS is not the key element for location-based services because of some issues where GPS is unfeasible in dense urban areas or inside buildings where satellites are not visible from the mobile terminals; they propose the position and motion tracking system (PMTS). PMTS differs from GPS in three respects: i) it exploits only information available to the network itself, ii) it does not require additional hardware and iii) it provides mobility prediction at both the network and the service level.

Some of the previous studies for mobile location systems focus on privacy concerns in using location-based services. Louise and Anind [11] found that, on a scale of 1 to 5, the participants in their study averaged 2.75 for privacy concerns. This study showed that people are not overly concerned about

[3] http://www.pc.pitt.edu/transportation/routes.php

their privacy when using location-based services. On the other hand, another study focuses on privacy concerns for location tracking services, which showed that 3 out of 16 participants were highly concerned and the rest were not concerned or didn't mind either way. The final conclusion of this study is that people are more concerned when others can track their location than when their phone reacts to its own location.

Information that can be taken from the GPS is highly needed in many applications. Tracking Pro is one of these applications. The GPS Phone Tracking Pro application makes it simple to track the movements of other mobile devices by allowing observers to find the exact location of the observed users and getting real-time updates about their movements. It is used most of the time to locate any phone that has been misplaced or stolen[4] . Another application is U Safe Tracker application[5] ; its ultimate purpose is the safety of the observed users by reporting their locations to the observers. This application uses the mechanism: Set the application to send the observed user's location every 30 minutes or at a specific time only (e.g. 9:00 PM every day).If no location signal is received after 30 minutes (or at a time specified by the observer), the phone's owner may be in danger.

Mobile Tracker[6] is also an application that keeps track of other mobile phones. Its ultimate goal is tracking lost mobile phones, and it works as follows:

- Create an account so observers can login from anywhere.
- Observers should pick four contacts by providing their mobile numbers.
- When this mobile gets lost or stolen and when the new SIM card is inserted in the mobile, this application automatically sends SMS from the SIM card that is inserted.
- Once the SMS is received, observers can make a complaint on that number that they received SMS from and the mobile will be tracked.

Mobile Number Tracker[7] is an application that provides information about calls that are received from unknown numbers and observers will be able to find out the location of the caller. The proposed developed application is enhanced version of the existing applications with more features that coordinate the work of three end users (transportation administrator, bus driver, parents and passenger (student)), they use similar techniques and tools but are differ in purpose.

This problem entails the need for mobile application that is able to better utilize the Global Positioning System (GPS) capabilities of a mobile phone and provide services that overcome the currently existing problems. This mobile application is developed for Imam University Transportation System and tested by transportation administrators, bus drivers, and students of Imam University, Riyadh, Saudi Arabia. It is compared with the existing applications and a comparison table in terms of tools and technologies is as shown in Table 1.

TABLE I. COMPARISON BETWEEN TECHNOLOGIES

	Systems				
	[4]	***[5]***	***3***	***[6]***	***Proposed System***
GPS	✓	✓	✓	✓	✓
GIS	✓	✗	✗	✗	✓
Maps	✓	✗	✓	✗	✓
Time precision	✓	✗	✗	✓	✓
SMS and GSM	✓	✓	✓	✗	✓
Energy Monitoring	✓	✗	✗	✗	✓
Internet Requirement	✓	✗	✗	✓	✓
Traffic Management	✓	✗	✗	✗	✓
Desktop and Web Application	✓	✗	✗	✗	✓
Smart Phone Application	✗	✗	✓	✓	✓

II. METHODOLOGY

The algorithm that Google adopts to find the shortest path is Dijkstra's algorithm. The idea of this algorithm is finding the shortest path from one specific node to another. This algorithm works on a weighted graph that has a start node and a goal node and the mission is finding the least cost path to the goal node. The Dijkstra's algorithm works as follows:

- In the beginning, the distance of each node will be set to ∞ with the exception of the first node that we will start from that is set to 0. In addition, each node will be marked as unprocessed.
- There will be visited nodes that start with the first node and unvisited nodes that start with the rest of the nodes.
- In the current node, find out all of its neighbors that are unvisited and calculate (the distance of the current node PLUS the distance from the current node to its neighbor). If the result of this addition is less than their current temporary distance, it should be replaced with this new value.
- The node is considered to be visited and removed from the unvisited set if we are done considering all of its neighbors.
- The algorithm will stop and considered to be finished when the goal node status is changed from unvisited to visited.
- The unvisited node that has the smallest temporary distance must be marked to be the "next" to the current node and then repeat the steps starting from step 3.

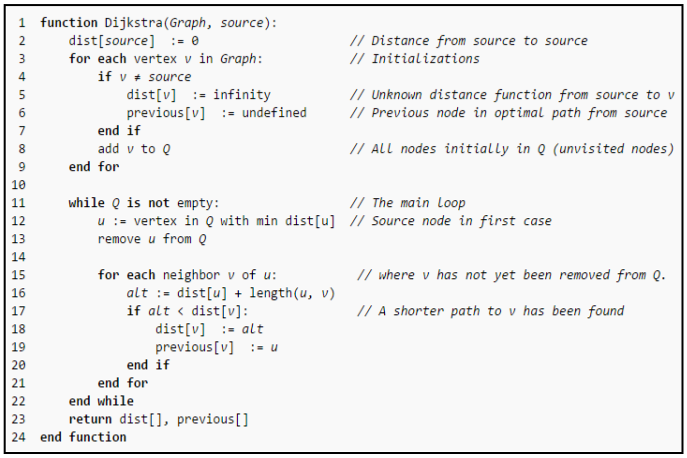
Fig. 1 shows the Dijkstra's pseudo code.

[4] https://play.google.com/store/apps/details?id=com.fsp.android.c&hl=en

[5] https://play.google.com/store/apps/details?id=homesoft.app.falcontracker&hl=en

[6] https://play.google.com/store/apps/details?id=com.nav.mobile.tracker

[7] https://play.google.com/store/apps/details?id=com.ViQ.Productivity.MobileNumberTracker&hl=en



```
function Dijkstra(Graph, source):
    dist[source]  := 0                     // Distance from source to source
    for each vertex v in Graph:            // Initializations
        if v ≠ source
            dist[v]  := infinity           // Unknown distance function from source to v
            previous[v]  := undefined      // Previous node in optimal path from source
        end if
        add v to Q                         // All nodes initially in Q (unvisited nodes)
    end for

    while Q is not empty:                  // The main loop
        u := vertex in Q with min dist[u]  // Source node in first case
        remove u from Q

        for each neighbor v of u:           // where v has not yet been removed from Q.
            alt := dist[u] + length(u, v)
            if alt < dist[v]:               // A shorter path to v has been found
                dist[v]  := alt
                previous[v]  := u
            end if
        end for
    end while
    return dist[], previous[]
end function
```

Fig. 1. Dijkstra's pseudo code

Three-tier system architecture is used for developing mobile application as shown in Fig. 2. The selection of this type of architecture is due to the following reasons. First, Three-tier architecture improves security that is needed in the application; the application layer will get the data after ensuring that a particular user has the right to access this data. Second, it facilitates in maintenance of the application. Third, it optimizes reusability.

- Data layer: It contains the local data.
- Application layer: The system uses this layer to apply the important functionality such as the operations related to logging, exception handling and validation.
- Presentation layer: It permits users to interact with our system through the user interface. It also has components that take the input from the user and validate it based on some constraints.

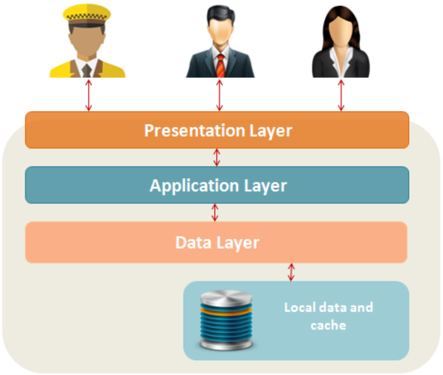


Fig. 2. System architecture

The functional requirements of the application are as shown in Fig. 3 with the help of use-case diagram.

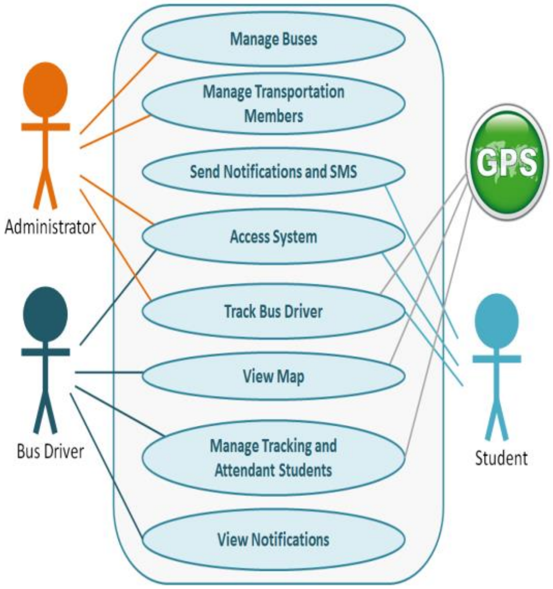


Fig. 3. Funtional requirements

III. Results

In this project a web application is developed using PHP and mysql database to be managed by the administrator on windows platform using Apache server with Intel i7 processor, 4 GB RAM and 512 MB hard disk. On the other end students and drivers use this information and communicate with each other using the JSON object of Android[8] operating system 2.2 or above after installing the .apk file in their mobiles with Quad Core 1.2 GHz processor, 1 GB RAM as minimum requirements. The developed application is tested with a sample of 30 students from Imam University, 3 buses operating on different routes handled by one administrator. Fig. 4 shows the few screen shots of web application used by the administrator. Fig. 4(a) shows the home page of the developed web application. Upon login to the system the buses related information can be viewed in buses tab as in Fig. 4(b). A new bus information can be added by filling the details related to the bus as shown in Fig. 4(c) and the confirmation message is viewed in Fig. 4(d). The transportation members tab has three options to select Fig. 4(e), upon selecting transportation administrators, the system allows to add, view, edit and delete services as shown in Fig. 4(f). Selecting bus drivers option Fig. 4(g) offers the services of add, view, edit and delete the information related to bus drivers. Similar services are offered by the system as shown in Fig. 4(j) for the student's option in Fig. 4(i). The details related to a particular student can be viewed upon clicking the view in students page Fig. 4(k). The students location on the map can be viewed as shown in Fig. 4(l) upon clicking the location on map in students page. Fig. 4(m) shows all the bus locations on the map upon selecting bus locations tab.

[8] http://www.businessinsider.com/12-ways-android-is-still-better-than-ios-7-2013-9?op=1

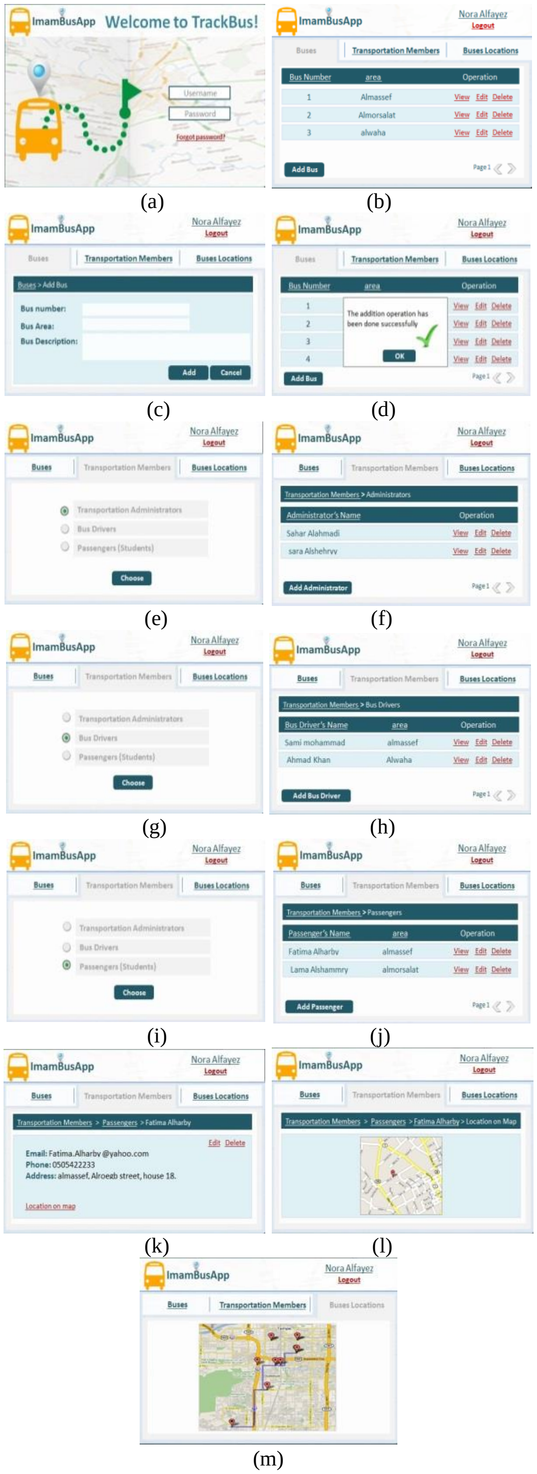


Fig. 4. Administrator functions

Fig. 5 shows the few snap shots after installing the .apk file in an Android mobile used by the drivers. Fig. 5(a) shows the login screen and after successfully logged on the driver view the shortest path to the destination as shown Fig. 5(b). The driver manage the busload by adding or deleting students from Fig. 5(c) upon selecting the Manage Bus Load option from Fig. 5(a). The driver can view the location sent by the students as shown in Fig. 5(e) from selecting the view option of a particular student in Fig. 5(d) upon selecting View Students Notification service in Fig. 5(a). The driver can send the message to any selected student from selecting contact as shown in Fig. 5(f). The driver has the privilege to change his password from settings as shown in Fig. 5(g) and can view details of the application in Fig. 5(h) from About service which is at the bottom of the screen.

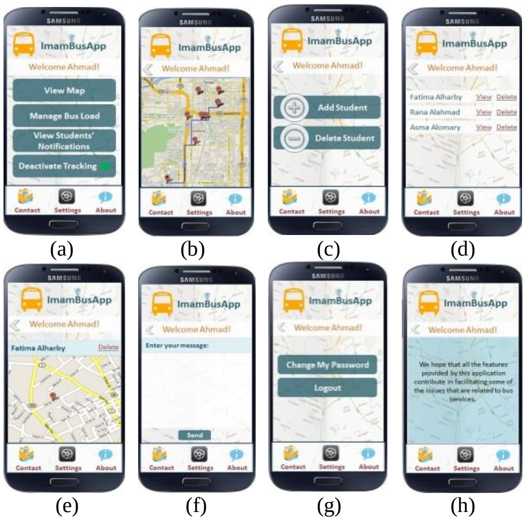


Fig. 5. Bus driver services

Fig. 6 shows few interfaces used by the students. Fig. 6(a) shows the home screen of the student upon successful login. The student can see two options, upon selecting Track My Driver in Fig. 6(a) The student can see the position, location of the driver on the map and the current bus load as shown in Fig. 6(c) and Fig. 6(d).

The students can inform the bus driver for their absence any time prior to the bus arrival to the bus-stop by selecting Absent Tomorrow service in Fig. 6(a) which can be reconfirmed by the system in Fig. 6(e) and message can be successfully delivered to the driver as shown in Fig. 6(f).

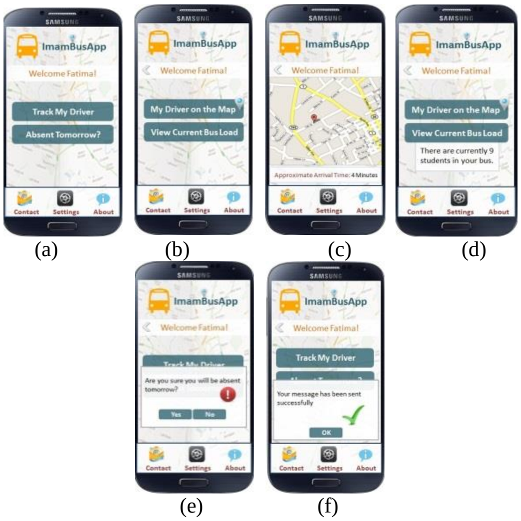


Fig. 6. Student services

IV. Conclusion

After testing the developed application a questionnaire is distributed to evaluate the students satisfaction. Very promising results with a satisfaction rate of 98.7% are achieved. This application not only helps transportation administrators of Imam University to manage transportation services that are related to buses and students efficiently and effectively but also serves vast segments of society. The developed application can be used by any transportation system and also enable parents to track their children, employers to track their employees, track the elderly people etc. It can also be utilized to register movements for the tracker himself/herself to go back and explore the places that they were in at any given time.

Enhancement to this system can be done as a part of future work by adding the following features.

- Having programmed hardware to act as a control unit that is attached to each bus. It is more convenient to track the bus vehicle itself instead of tracking the bus driver.
- Providing a more improved way of informing the student about getting into her area. Once the bus enters the area of a particular student, a notification is automatically sent or an SMS message
- Having sensors in the bus whereby, each time someone gets in, the counter of the number of the students on the bus increases. Also each time someone gets out, the counter decreases.
- Having a chat room for the students who live in the same area and who share the same bus.
- Showing the students the route of the day on a map and to give an idea of student numbers on any specific day. If many students are coming, this means the route is longer and if few students are coming, the route is shorter. Having an idea about how long the route of the day would allow the students to know when the bus may come and pick them up, thus saving them from having to wait for an unknown length of time.

Because of the fact that there are two subsystems that are collaborating with each other (the web application and the Android application) to produce the necessary information, the system's performance is not very high. This is one of the limitations of this system and finding techniques to improve the performance of the system can also be a part of the future work.

References

[1] R. Meier, “Professional Android 4 Application Development,” John Wiley & Sons, 2012.

[2] D. Huber, “Background positioning for mobile devices - android vs. iphone,” In Joint Conference of IEEE Computer & Communication Societies, 2011.

[3] E. Oliver, “A survey of platforms for mobile networks research,” ACM SIGMOBILE Mobile Computing and Communications Review, vol.12, pp 56–63, October 2008.

[4] O. O. Emmanuel and M. N. Moses, “GIS based public bus transport management system for Nairobi city,” 1 st Esri Eastern Africa User Conference (EAUC), Nairobi, Kenya, 17–18 September, 2013.

[5] A. Kannaki, N. Vijayalashmy, V. Yamuna , G. Rupavani and G.Jeyalakshmy,“GNSS based bus monitoring and sending SMS to the passengers,” International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, Special Issue 1, pp. 2502–2506, March 2014.

[6] M. E. Mallia and K. Simpson, “Wireless global positioning system fleet tracking system at the university at Albany. Report No. C-11-12/ 14-27, 2014.

[7] N. Marmasse and C. Schmandt, “Save and sound : a wireless leash,” In Proceedings of CHI '03 Extended Abstracts on Human Factors in Computing Systems, Ft. Lauderdale, FL, USA, April 05–10, 2003, pp. 726–727.

[8] S. Motahari, H. Zang, S. Bali, and P. Reuther, “Mobile applications tracking wireless user location,” In Proceedings of IEEE Global Communications Conference (GLOBECOM), 3–7 Dec. 2012, Anaheim, CA, pp. 2006–2011.

[9] H. Lee, I. Park, and K. Hong, “Design and implementation of a mobile devices-based real-time location tracking,” In The Second International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Sept. 29 2008–Oct. 4 2008, Valencia, pp. 178–183.

[10] M. Anisetti, V. Bellandi, E. Damiani, and S. Reale. “Advanced localization of mobile terminal,” In International Symposium on Communications and Information Technologies, 17–19 Oct. 2007, Sydney, NSW, pp. 1071–1076.

[11] L. Barkuus and A. Dey, “Location based services for mobile telephony: a study of users' privacy concerns,” In Proceedings of Interact 2003, Zurich, Switzerland, 2003, pp. 709–712.

Adoption of e-Government in Pakistan: Supply Perspective

Zulfiqar Haider[1] , Chen Shuwen[2]
Dalian University of Technology,
Faculty of Management and Economics,
No.2 Linggong Road, Ganjingzi District,
Dalian 116023, Liaoning, P.R. China

Dr. Farah Lalani[3] , Dr. Aftab Ahmed Mangi[4]
Department of Public Administration
Faculty of Social Sciences,
University of Sindh, Jamshoro
71000, Sindh, Pakistan

***Abstract*—Electronic Government, also known as e-Government, is a convenient way for citizens to access e- services and to conduct business with the government using the Internet. It saves citizens and the government both time and money. This study examined adoption of e-Government supply side by looking at the UTAUT as a model of technology acceptance. Furthermore, specific variables that were proposed to moderate relationships within the UTAUT were analyzed including locus of control, perceived organizational support, affective and normative commitment, and procedural justice. Data from one sample indicated that in general, the UTAUT model was supported, however, the moderators proved non-significant. Implications are discussed for the technology acceptance process as technologies are implemented within countries and suggestions for future research in this area are discussed. This research sought to demonstrate the robustness of trust-based UTAUT to address e-Government adoption concerns. As a consequence, it was the responsibility of the researcher to select research questions, operational variables, research approaches, and research techniques, within the scope of the study. The research hypotheses formulated in this study were based on the technology acceptance literature covering the original UTAUT model with the inclusion of the trust construct. This quantitative study was conducted with help of Unified Theory of Acceptance and Use of Technology (UTAUT) model.**

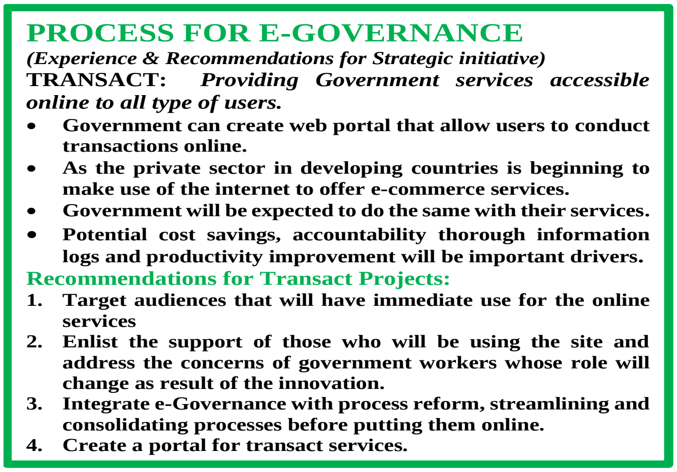
***Keywords*—e-Government; adoption; Supply; UTAUT model; Pakistan**

I. INTRODUCTION

The implementation and use of technology in government agencies, in order to improve access to information, procedures and services offered to citizens; increase the efficiency and effectiveness of public administration and to substantially increase transparency and citizen participation are the main concepts related to electronic government. The implementation of electronic government should be viewed as the right of citizens to interact electronically with government. This implies that administrations are interrelated to simplify procedures, services and procedures.

At present the use of new technologies is fundamental to support the transformation processes that are being developed tool. [1]. The Electronic Government includes all initiatives that enable the migration of information (procedures and services of paper-based manual procedures) of computerized procedures; accessing them through multiple channels such as the Internet, mobile devices, Citizen Service Centers, among others [2]

Before analyzing the role played by the background of UTAUT, begin by exposing the resulting relationships between the component variables of the model used, verifying the impact of expectancy and effort expectancy result on the intended use of the platform e-Government. According to the results achieved when a person has freedom in the decision to adopt e-Government platform it will if considered useful and easy to use; also in line with social influence plays no significant role in the adoption of e-Government. As suggested in the literature, in the case of voluntary adoption environments, attitudes correlate with behavioral intention and using this according to the results achieved and in line with previous studies. According to UTUAT model, it is also important particular for public sector employees have a professional skills with positive behavior to provide best e-Government services to their citizens [3]. Even though earlier researchers missed to identify to know the important factors related to supply side development. But it is crucial to analyze the factors that implementing the adoption of supply-side of e-Government in developing countries like Pakistan.



PROCESS FOR E-GOVERNANCE
(Experience & Recommendations for Strategic initiative)
TRANSACT: ***Providing Government services accessible online to all type of users.***

- **Government can create web portal that allow users to conduct transactions online.**
- **As the private sector in developing countries is beginning to make use of the internet to offer e-commerce services.**
- **Government will be expected to do the same with their services.**
- **Potential cost savings, accountability thorough information logs and productivity improvement will be important drivers.**

Recommendations for Transact Projects:

1. **Target audiences that will have immediate use for the online services**
2. **Enlist the support of those who will be using the site and address the concerns of government workers whose role will change as result of the innovation.**
3. **Integrate e-Governance with process reform, streamlining and consolidating processes before putting them online.**
4. **Create a portal for transact services.**

Fig. 1. e-Government Pakistan (fbr.gov.pk)

Governments are making a major effort to develop systems providing public services that come to complete the traditional. The effective use of the platform will be caused both by the intention of using it as the belief that there have both human resources infrastructure and personnel needed to help if necessary [4].

The Electronic Government, in its broadest sense, is developed mainly based on four categories of stakeholders:

- Citizens and organizations or civic associations.
- The private sector, through individual and corporate operators.
- The state, through state employees and / or other government and state agents.

This work aims to fill this gap by presenting a more holistic view of e-Government development in the municipalities of the metropolitan area of Sindh using this index for telecommunications infrastructure households, the index of the cultural capital of the population and the rate evolution of the websites of the municipalities of the metropolitan area; in this way you will know if the benefits of e-Government reach most of the citizens of these municipalities.

II. Theoretical Foundation

The obstructions are abridged in the lack of involvement in the utilization of IT, lack of mindfulness and information of e-Government administrations and lack of trust in government and IT alike. Utilizing the UTAUT, investigated the supply side factors that influence the use of citizen driven e-Government administrations in Pakistan [5]. The outcomes demonstrated that factors, for example, encouraging conditions, companion impact, execution hope, and exertion anticipation, clarify the conduct of Pakistani citizens towards the use of such e-administrations. The study additionally indicated different factors, for example, society, trust, which was excluded in this connection, yet in future studies in light of their significance.

Also in distinctive connection, investigated utilizing the UTAUT the supply-side factors that influence the behavioral expectation of people to use citizen driven and e-Government administrations in Pakistan [6]. The outcomes demonstrated three compelling factors, execution anticipation, social impact, and encouraging conditions [7]. The study likewise indicated different factors, for example, society, which was excluded in this connection, yet in future studies in view of its significance. Talked about the factors influencing the citizens' plan to adopt and utilization citizen-driven e-Government administrations, especially the difficulties confronting the Information Technology move in the Pakistan on a substantial scale. The outcomes demonstrated that among those factors, nature of administration, dispersion of advancement, PC and information literacy, society, lack of mindfulness, specialized base, site outline, and security, which influence the citizens to adopt e-government administrations in the same setting [8].

The e-Government initiatives aimed at rural and marginal urban areas, as well as groups that have traditionally been disadvantaged, such as indigenous peoples and women, should aim at improving their quality of life and work. They should also aim to reduce poverty by encouraging participation in political processes, the design of effective mechanisms to address the most pressing needs and opening spaces for insertion in the labor and productive sphere. Many public services are offered online for citizens to use. Citizens are now able to pay their property and income taxes online, file their tax returns, apply for unemployment services, renew their drivers' licenses, research political candidates or elected officials and their platforms, file complaints, register to vote, and participate in electronic voting [9]

In Pakistan citizens can also search public websites and library websites, change their addresses, register vehicles, file police reports, apply for jobs, and download official forms on the Internet. According to Asgarkhani (2007), the benefits of e-Government include providing expedient government services to citizens and businesses, improving the economy, allowing for greater public access to information, empowering citizens, and making the government more accountable to its citizens. Yet, these benefits are not equally distributed [10].

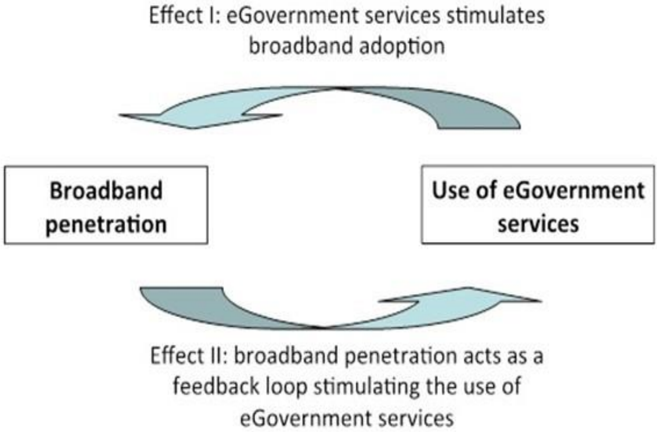

Effect I: eGovernment services stimulates broadband adoption
Broadband penetration
Use of eGovernment services
Effect II: broadband penetration acts as a feedback loop stimulating the use of eGovernment services

Fig. 2. e-Government services broadband adoption

Official efforts for the establishment of e-Government in Pakistan started in the year of 2003 along with the establishment of an organization named "Pakistan Digital" by the government. Pakistan Digital is liable for all e-commerce and electronic government services in the country. From the get go, this association was accountable for identifying the technological and information requirements for different agencies of government in Pakistan in order to contribute in electronic government [11] [12]. According to the recent study, the electronic government of Pakistan is still in the beginning phase of building e-services that concentrate on providing data to users.

In Pakistan many government services are provided online and each year it appears that new services are added. According to Aerschot & Rodousakis, 2008 Common government services provided online include: downloading forms, obtaining information, using job services, searching for library books, submitting forms, making payments to different government entities, interacting with the tax office, requesting documents, completing change of address forms, statements to the police, and car registration. Citizens may also pay parking tickets; request passports, drivers' licenses, and birth certificates; and request other replacement documents online. Many of these service tasks can be accomplished in minutes with several key strokes [13].

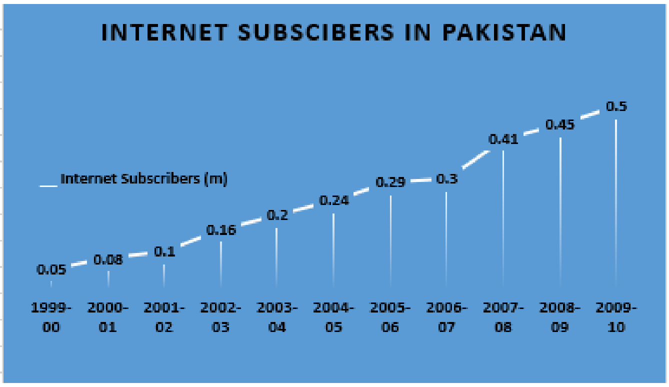


Fig. 3. MS Excel (output) - Source PTA (Pakistan Telecommunication Authority

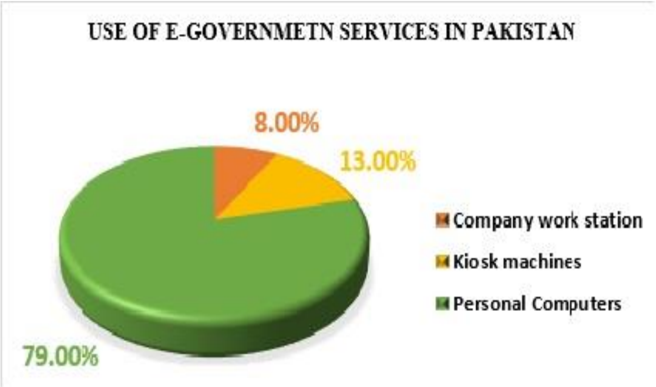


Fig. 4. use of e-Government services in Pakistan

The paper also verifies the significant effect of knowledge on social influence, and self-efficacy and assistance on enabling conditions, but the convenience of the latter. Nor is the effect of self-efficacy on the expectation of effort and convenience of enabling conditions observed. Being self-efficacy perceived the key cognitive effort, history can understand that this is not related significantly with the expectation of effort. Regarding the convenience that it is a voluntary service causes this is linked to the expectation of effort but not the enabling conditions. The services and facilities provided by the e-Government relate to the idea that benefit from its use will be easy, but not there technology and support required if needed citizens.

The economic and social crisis in much of Pakistan continually question the current organization and structure of public administration in the supposed interest of improving our competitiveness [14]. To achieve this, governments are taking measures of various kinds. On the one hand, they are delegating some of its functions to private entities (taxes and fees charges by banks and savings banks, for example); on the other, are trying to arrange with the public are made increasingly without resorting to personal interaction with the administration, thereby promoting self as usual in other economic sectors.

III. Literature Related to E-Services Adoption

In Pakistan, citizens contact the government for personal concerns and problems, and they also contact the government to influence public policy, and to locate information about government services and benefits (Cohen, 2006) [15]. The literature also points out that many of the services offered online do not allow citizens to conduct entire transactions or complete their tasks from beginning to end (Rodousakis & Santos, 2008). Citizens are forced to mail documents or go in person to a government agency when they would prefer to conduct their business online [16]. Possible reasons that entire transactions cannot be conducted online are the need to prevent fraudulent activity by identifying those applying for services, the funds are not available to make government websites fully functional or capable of completing transactions online, and to keep government employees employed [17] [18].

A good deal of research is available that has investigated the determinants for accepting information technology services in offline and online environments. Evidence from the literature shows that there are two main approaches for Information Technology acceptance research (Harrison et al., 1997; Hernandez and Mazzon, 2007; Taylor and Todd, 1995a). One of the most widely used approaches works on developing strategies to examine IT adoption. This approach suggests and uses models and behavioral theories that are drawn from psychology, which act as a foundation for information systems research, such as, the Technology Acceptance Model (TAM) and Theory of Planned Behavior (TPB) (Harrison et al., 1997). The theory of planned behavior, for instance, is famously established and is frequently used in various settings for research into IT adoption to determine intention behavior [19].

TABLE I. Frequently Used in Various Settings for Research

	Cities with websites (n=2376)		Small cities with websites (n=725)	
deviation	Mean	St. deviation	Mean	St.
Area 2000	**1184.22**	**2769.84**	**1349**	**6187.51**
Population 2006	**114685**	**348657.26**	**33002**	**100519.73**
Population 2005	**113400**	**346021.19**	**33230**	**101248.95**
Population 2000	**107195**	**329418.83**	**32837**	**99767.50**
Population 1990	**94013**	**296877.19**	**31121**	**95364.19**
Population density per Square mile	**273.67**	**1934.04**	**63.61**	**184.99**
No. of households	**40061**	**117041.83**	**12643**	**38260.89**

If e-Government or information technology is to facilitate an organizational "transformation" in the public sector, then it will also require a more critical examination of the way we are currently measuring such initiatives. Public agencies will need to expand their thinking about e-Government, and incorporate initiatives that build IT capacity into organizations. Finally, public organizations will need to move beyond the current e-commerce model and begin to look at how the Internet can be used between citizens, the business community, and internally between government agencies. Small cities tend to be challenged in the digital domain to deliver efficient services to their citizens, realize the potential of information and communications technology (ICT), and grant citizens large participatory roles in their governance [20]. E-government activity should bridge the communication gap between elected officials and stakeholders both more diverse and more extensive [21]. Barriers in supply side of e-Government in Pakistan were observed from a survey result.

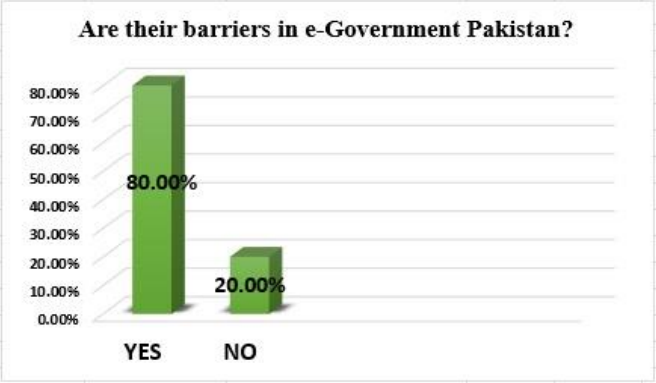


Fig. 5. Barrier in e-Government

Availability of e-Government services as part of an e-government 2.0 program had a 28% usage average, the lowest of the three benchmarks. Future research could be required as the availability of e-Government services as part of e-government 2.0 program increases [22]. Research by local government Information Technology professionals in areas such as rich site summary (RSS) as a means of feeding posted information through newsfeeds and e-mail programs and improving access to city services from mobile devices is becoming as it becomes more common. While all of the cities explored in this study had an official presence on the World Wide Web, only 10 of those cities optimized the websites for viewing on mobile devices. All of the cities which utilized third-party website development were optimized for mobile viewing [23]. A recommendation would to add mobile device viewing optimization as part of a third-party e-Government program development. And finally the a return on investment analysis by elected local government officials on using kiosks for access to e-Government services from locations conveniently located throughout the city. A major obstacle could be the cost of kiosks. A recommendation for a kiosk solution could be following the example of Farmington, Sindh's kiosk stations. Farmington purchased 21inch Android touch screen tablets with Site Kiosk software installed. The kiosk provides access to the e-Government services available on the city's website. Total cost of ownership was less than $400 per location [24].

Farmington has 4 kiosk locations at public buildings throughout the city. 147 Further researches on whether this is a viable solution for other small-sized rural cities could be conducted. [25]. The determination of this qualitative study was to identify, describe, and analyze user adoption patterns of e-government services in small-sized cities. To fulfill the effort of the study, this research utilized a multiple–case study design to examine user adoption patterns that remain undocumented in fourth class, third class, and constitutional charter cities in Sindh, Pakistan. The goal of this study was fulfilled through content analysis of the cities' official websites and social media sites. The results of this study suggest that while e-government services for small-sized rural cities still lag behind larger city and state government agencies, the local government agencies utilizing companies that specialize in the development of government to citizen applications for the development of their e-government services were scored the highest for the number of e-Government services offered.

The government agency websites developed by those companies were easier to navigate and offered the most services. Additionally, the sites were accessible and optimized for use on portable devices. Additionally, the results of the current study support Baumgarten and Chui's (2010) assertions that Web 2.0 tools should be incorporated with e-government tools to provide e-Government services that promote citizen empowerment and promote collaboration between tax payers and government agencies. However, the results of this current study showed that the digital divide characteristic of limited access to high-speed broadband Internet service was not a factor for the cities explored. With readily available access to high-speed broadband Internet, the question was raised as to why e-government services were not being offered consistently across the cities explored in this study.

This supports Christopher's (2009) assertion that the greatest barrier to delivering e-government 2.0 services is not just a factor of the digital divide but also a factor of cost and limited government budgets. It would seem that the cost of implementing an effective e-government 2.0 program with e-government services which empower citizens could be factor as to why adoption patterns is low for small-sized rural government agencies [26] Again this becomes apparent in the results of this study that show the cities that utilized third-party e-government program implementation have the highest number of services available to their citizens. Consistent with research literature which suggests that social media venues can be used to create lines of communication and interaction between government agencies and citizens (Auer, 2011; Bryer, 2011; Dadashzadeh, 2010; Hand & Ching, 2011; Mann, 2010) depending on rates of user adoption (Colesca, 2009), this research showed that the cities which utilized social media sites had interaction and collaboration with citizens. Lin and Lu's (2011) research showed that having an understanding of the citizens' 149 willingness to accept social media could enable elected government officials to develop social media sites that target the needs of the users. Of the cities explored in this study, 13 utilized some form of social media. In some cases, such as Farmington, the people following the social media sites outnumbered the total population of the city. This shows that if the city posts information that is relevant to the followers, the citizens of that city and surrounding areas will not only follow that information but also become participants by commenting and asking questions on the sites. This is the kind of collaboration that could be used to empower citizens in the democratic process. There may be multiple audiences that can benefit from this research. For example, the findings of this study may be of interest to local government elected officials, government agency departments, government IT professionals, and city citizens. Local government officials, government IT professionals, and government agency departments can use this study as a guide as to which services are being offered by other cities and which tools work best for empowering citizens. City residents could use this to see which services could be made available which could better serve the general public and empower citizens as an active participant in local government decision making processes. Overall, this study extends the body of knowledge on citizen participation through e-government services and could provide supporting data for leaders in local small-sized government agencies to advocate for and implement government transparency and citizen empowerment.

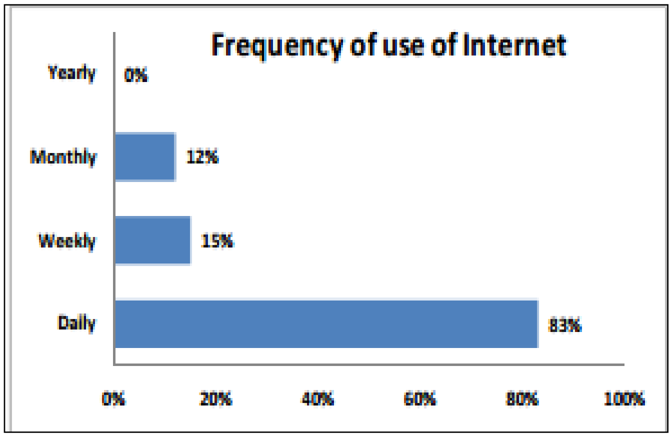


Fig. 6. Frequency of use for internet

IV. Purpose of the Study

The purpose of this qualitative study was to identify, describe, and analyze user adoption patterns of e-Government services in small-sized cities. To fulfill the purpose of the study, this research utilized a multiple case study design to examine user adoption patterns that remained undocumented in fourth class, third class, and constitutional charter cities in Pakistan with populations between 12,000 and 20,000 citizens [27].

The goal of this study was fulfilled through content analysis of the city's official websites and social media sites. An exploratory qualitative research approach was an appropriate method for exploring the purpose of the study because of a lack of confirmed and well established e-government services and user adoption patterns in the population (Shank, 2006). Quantitative study for internet access in Pakistan showed the following results.

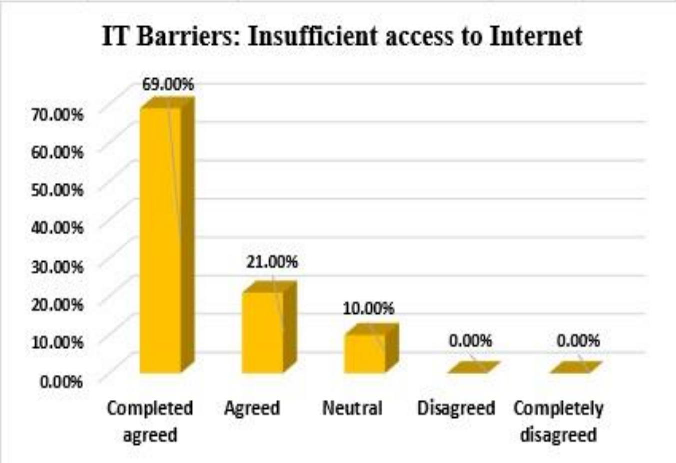


Fig. 7. I T Barriers

V. Research Question

The objective of first research question

1) Identify user adoption patterns of e-Government services in small-sized cities in Sindh. The objective of second research question.

2) Identify what a content analysis of user adoption patterns of e-Government services indicate about citizen empowerment opportunities in small-sized cities in Sindh.

Data was collected by analyzing user adoption patterns of the researched cities social media sites and available citizen satisfaction surveys. To evaluate the findings of this research, each city that made up the population of the study was treated as a separate case study. Since each of the sample cities were treated as a separate case study, an evaluation of the findings focusing on the 10 semi-structured based on the three benchmarks of e-participation established by the United Nations 2010 survey (United Nations, 2010) are listed below. Fulton. Fulton offers some of the e-Government services expected from a small sized city. The official website was difficult to navigate when looking for specific documents to download.

VI. Significance of the Study

The Adoption of e-Government in Pakistan is important in empowering citizens and making them an integral part of the democratic process. There was evidence that technological advances in e-Government, especially ICTs, have been utilized to enhance dissemination of information and citizen participation in government processes (e-tax system from Federal Board of Revenue (FBR). However, the body of evidence on its adoption and usage by citizens remained weak. Developing an in-depth understanding of the adoption patterns of e-Government in small-sized rural cities in Pakistan was the goal of this study. Findings from this study may be useful for promoting citizen empowerment within small-sized rural cities. This goal was accomplished by understanding the barriers to empowerment through e-Government is presented and making recommendations on how to overcome those barriers to user adoption of e-Government in small-sized cities in Pakistan.

Regarding the operation UTAUT model the expected result (path = 0.3102, t-value = 2.7378) and the expectation of effort (path = 0.4204, t-value = 4.2514) directly affect the intended use, while social influence no such significant effect (path = 0.0531, t-value = 0.5696). Finally, both the intended use (path = 0.2719, t-value = 3.2215) as enabling conditions (path = 0.2118, t-value = 2.3349) significantly affect the effective use of the platform of e-Government.

VII. Implications, Recommendations, and Conclusions

The basis of this study was an examination of user adoption patterns of e-Government services in small-sized cities in Sindh- Pakistan. The problems addressed were a gap in identifying the user adoption patterns of e-government services and what a content analysis of user adoption patterns of e-Government services indicated about citizen empowerment opportunities. Despite the policy calling for government transparency, citizen interaction, and government openness through the use of e-Government 2.0, little has been done in this area for small-sized cities [28]. Particularly those with limited access to high–speed Internet service [29]. This phenomenon was particularly true for e-Government, the one section of e-Government with the lowest adoption rate [30] and the greatest potential for citizen empowerment. The purpose of this qualitative multiple-case study was to identify, describe, and analyze user adoption patterns of e-Government services in fourth class, third class, and constitutional charter cities in Sindh, Pakistan.

For this research paper, a qualitative approach with a multiple case study design was used to answer the research questions. Due to the limited availability of relevant research data on e-Government and e-government initiatives for small-sized rural cities [31]. According to Akbulut-Bailey, 2011; Bonsón, Torres, Royo, & Flores, 2012; Parvez, 2008, this study was exploratory in nature in order to uncover relevant information about the topic. Exploratory research was a viable option for the research of citizen empowerment through e-Government services (Zikmund, Babin, Carr, Griffin, 2010). This study is non 139 experimental in nature investigated the challenges, practices, and strategies related to the delivery of e-government services as observed in the 23 small-sized cities and 4 census designated places of Sindh that are the focus. Data collection included relevant and available sources such as documentation and observations from the city's official website and social media presence. Data collection focused on 10 semi-structured questions that were used to identify and measure e-government services offered by small-sized Sindh cities which were the focus of this study.

According to our research, if governments want to increase both the intent and the effective use of e-Government in the Administration should develop actions of public marketing to improve the perceptions that people have about their expectations of use, on its ease of use and on the existence of willing and able to solve any problems that arise during use professional support. The work presented confirms the relevance of these elements from the perspective of managed also highlighting the importance for management is considering these 4 factors: aversion citizen expressing the personal interaction with the Administration, the degree of confidence that holds in the tool, receive assistance and convenience you see in the use of e-government.

Consequently, aversion to personal interaction is a key segmentation criteria that should be used by management to identify population groups of citizens who develop a different behavior in relation to the use of e-Government. It is recommended, therefore, to distinguish groups in the population according to the degree of aversion (high or low). In both segments is advisable to implement actions that enhance communication with messages on the citizen trust in e-Government tools and security in obtaining assistance. However, since the motivation for use in population with low aversion to personal interaction ventures to lower priori, the effort it must also focus on showing convenience (less effort and better conditions of service provision).

A. *Implications*

Two research questions guided the purpose of this study.

The findings represent a contribution towards a better understanding of how small-sized local government agencies in Sindh are empowering citizens through e-Government services. Consistent with the purpose of this study, Research Question One (Q1) asked: what are the user adoption patterns of e-Government services in small-sized cities in Sindh? Research question two (Q2) identified that cities which contracted third-party e-Government software developers to create an e-Government 2.0 portal which incorporates e-Government tools for user collaboration, communication, and participation have the highest user adoption rates.

Additionally, e-Government services are more likely to be adopted by cities, which have a social media presence that promotes citizen participation [32]. The findings indicate that adoption patterns of e-Government services in small-sized cities, from most prevalent to least are, are based upon availability of e-Government services, communication between government agencies and citizens, and government transparency. A finding of this study is that cities, which outsourced website development with integrated e-government and e-Government tools, have higher user adoption and availability. D'Agostino's et al. (2011), Park (2007), and all agree that static websites that are typical of small-sized local government agencies are ineffective in delivering the e-Government services expected by citizens in the digital age. In fact, the rate of adoption of ICTs to support citizen empowerment in the form of e-Government was assessed as "fair" to "low" in the extant [33, 34].

The implication of this current research may indicate that 141 if small-sized local government agencies utilize the services of companies that specialize in the development of government to citizen applications for the development of their e-Government services, they could improve the availability of services that could empower citizens. One interesting finding of the study is that within the past 5 years high-speed broadband Internet service has become readily available in the cities explored in the study. Baird, Zelin, and Booker (2012) found that there seems to be an increasing disparity in the development and adoption of e-Government programs, especially in small-sized cities, due to the digital divide. The implication of this current research may indicate that one of the primary factors that was found to increase the digital divide gap, lack of high-speed broadband Internet, did not seem to be a factor in the cities and census designated places explored in the study.

Every one of the 23 cities, information is unknown on the census designated places, provided access to high-speed broadband Internet in the Public Libraries as well as other public accessible buildings. Past research has demonstrated that even though the Internet offers open access to political information and services, this primarily benefits those with easy access to the high-speed Internet service (Schwester, 2009).

For example, Whitacre (2010) indicated that rural small-sized communities are at a political disadvantage and cannot take advantage of modern ICTs for e-government services due to lack of accessible and affordable broadband Internet service. Additionally, Alemanne et al. (2011) stated that the rural community Public Library should be the leader in broadband high-speed Internet access for the location they serve [34], [35].

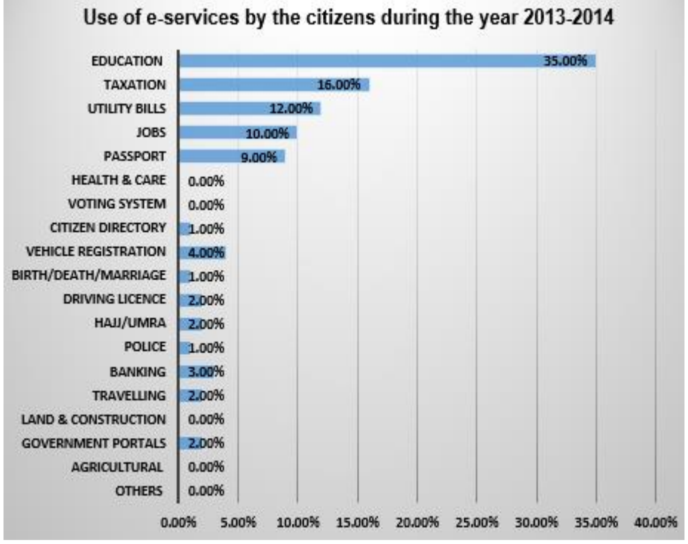


Fig. 8. Use of e-services by the users

The Taylor et al. (2012) indicated that Public Libraries which offer access to high-speed Internet not only improve availability of e-Government 142 services, but they also often have staff available that can help train the user on how to best utilize those services. The cities explored in this study all have broadband Internet service with public access computers at their local Public Library and Gautam et al. (2013), explained that the few remaining areas without broadband Internet service are currently constructing broadband infrastructures through the broad band Now grants. Research Question Two (Q2) asked: what does a content analysis of user adoption patterns of e-Government services indicate about citizen empowerment opportunities in small-sized cities in Sindh? Research question two (Q2) identified that the greatest citizen empowerment opportunity identified was the use of social media. While social media sites do take some time to create and manage, they are free tools that provide citizens with a means of two-way communication when used correctly.

The citizens become an active participant in the democratic processes of the local government agency. Questions can be asked and feedback provided in real-time. Additionally, social media can be used to disseminate news and events, take surveys, and solicit public input. Research showed that social media venues can be used to create lines of communication and interaction between government agencies and citizens. Depending on rates of user adoption (Colesca, 2009). The implication of this current research helps confirm the existing research theory that social media does increase communication between the government agencies and citizens, based on the observations of the cities explored in this study [36], [37] researched the challenges and obstacles that may be faced when government agencies implement social media. However, based on the findings of this study, social media is the preferred method of online communication between 144 citizens and government officials. The development of e-Government must be assumed as an evolutionary process into five phases (presence, interaction, transaction processing and citizen participation) and must meet four dimensions (external, promotion, internal and relational). These phases and dimensions are not interdependent nor need to complete one to start another. Each has a different purpose and requires different requirements in terms of organization, costs, needs knowledge and level of ICT use.

Although the implementation of e-government requires the availability of a technological infrastructure, this alone does not achieve the success of the transformation. To do human resources are required mastery of concepts of e-government, given their potential and level of technological literacy and during deployment managers at various levels of government develop a high motivation for change and achieve leverage advantages of e-government and minimize, during implementation, possible disadvantages that may arise.

B. Recommendations

The nature of exploratory research is that it raises more questions than answered and is often conducted to define research questions (Yin, 2009). Exploratory qualitative research approach was an appropriate method for exploring the purpose of the study because of a lack of confirmed and well established e-Government services and user adoption patterns in the sample population [38]. In this multi-case study qualitative research project, some questions were answered while others were partially answered and provide opportunities for further research. In addition to extending the body of knowledge about e-Government 2.0 programs and citizen empowerment through 144 e-Governments in small-sized rural Sindh cities, there are practical applications of this study in promoting citizen collaboration and participation. This section will address the recommendations for practice. The recommendations drawn from this study are relevant to elected local government officials, local government IT professionals, and the citizens of small-sized cities.

We have talked about a few obstructions in the development and movement of e-government. Keeping these factors aside, hierarchical and political factors are also, will remain the primary hindrances towards this new and better government at any rate in the creating countries. The governments who have the capacity to handle the red tape and political weight will be ensured pioneers in this type of administration. In future explores, the model proposed in this paper can be operational to limited down demand and supply side obstructions to add to a suitable system to further advance this manifestation of administration. Besides, more nation particular studies can be led to further enhance the proposed model.

C. Conclusion

The implementation of e-Government has been proven to be a convenient and effective way to boost productivity of government agencies and empower metropolitan area citizens by improving communication and promoting government transparency. However, the adoption of e-government services by citizens is inconsistent, at best, across the federal, state, and city government agencies. Visiting websites of smaller cities, it was clear to see that their e-Government initiatives consists of little more than just a static website with little to no online e-government services. Elected officials of small-sized rural cities operate under pressure from internal and external stakeholders to not only create government transparency through e-government initiatives but to also increase user adoption of the programs that are part of e-Government initiatives. This multiple case study qualitative study described how municipalities use present e-government adoption patterns in small-sized cities to generate user adoption and citizen empowerment.

The research questions were generated based on a comprehensive literature review and with the purpose of defining the approach for the present study. Consistent with the purpose of this study, the two guiding research questions were developed. A qualitative multiple–case study design was chosen because of the depth of understanding of the phenomena under study afforded to the researcher. This research study provided a better understanding of how local government 18 elected officials would use present e-government adoption patterns in small-sized cities to raise user adoption and to raise citizen empowerment.

Availability of e-democracy services as part of an e-government 2.0 program had a 28% usage average, the lowest of the three benchmarks. Future research could be required as the availability of e-democracy services as part of an e-government 2.0 program increases. All of the cities which utilized third-party website development were optimized for mobile viewing. This supports Christopher's (2009) assertion that the greatest barrier to delivering e-Government services is not just a factor of the digital divide but also a factor of cost and limited government budgets. There may be multiple audiences that can benefit from this research.

References

[1] Layne K, Lee J. Developing fully functional Egovernment: A four stage model [J]. Government information quarterly, 2001, 18(2): 122-36.

[2] Bokhari H, Khan M. Digitisation of electoral rolls: analysis of a multi-agency e-government project in Pakistan; proceedings of the Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, F, 2012 [C]. ACM.

[3] Lammers R. The Adoption of Open Standard Inter Organizational Systems [D]; University of Twente, 2010.

[4] Fang Z. E-government in digital era: concept, practice, and development [J]. International journal of the Computer, the Internet and management, 2002, 10(2): 1-22.

[5] Haider Z, Shuwen C, Panhwar K N. Demand for e-Government Adoption in Pakistan [J]. Public Policy and Administration Research, 2015, 5(5): 115-31.

[6] Ahmad M O, Markkula J, Oivo M. Factors influencing the adoption of e-government services in Pakistan; proceedings of the Proceedings of the 9th European, Mediterranean & Middle Eastern Conference on Information Systems, F, 2012 [C].

[7] Norris D F, Reddick C G. Local e-government in the United States: transformation or incremental change? [J]. Public Administration Review, 2013, 73(1): 165-75.

[8] Rehman M, Esichaikul V. Factors influencing the adoption of e-government in Pakistan; proceedings of the E-Business and E-Government (ICEE), 2011 International Conference on, F, 2011 [C]. IEEE.

[9] Bird R M, Zolt E M. Technology and Taxation in Developing Countries: From Hand to Mouse [J]. Nat Tax J, 2008, 791-821.

[10] Starling G. Managing the public sector [M]. Cengage Learning, 2010.

[11] Wescott C G. E-Government in the Asia-pacific region [J]. Asian Journal of Political Science, 2001, 9(2): 1-24.

[12] Asgarkhani M. The Effectiveness of E-Service in Local Government: A Case [J]. Asymptotic and Computational Methods in Spatial Statistics, 2009, 22.

[13] Torres L, Pina V, Acerete B. E-government developments on delivering public services among EU cities [J]. Government Information Quarterly, 2005, 22(2): 217-38.

[14] Herani G M, Lodhi S A. Innovative technology, social and economic sustainability: Evidence from Pakistan [J]. 2008,

[15] Lean O K, Zailani S, Ramayah T, et al. Factors influencing intention to use e-government services among citizens in Malaysia [J]. International Journal of Information Management, 2009, 29(6): 458-75.

[16] Alruwaie M, El-Haddadeh R, Weerakkody V. A Framework for Evaluating Citizens' Expectations and Satisfaction toward Continued Intention to Use E-Government Services [M]. Electronic Government. Springer. 2012: 273-86.

[17] Gupta M, Jana D. E-government evaluation: A framework and case study [J]. Government information quarterly, 2003, 20(4): 365-87.

[18] Ebrahim Z, Irani Z. E-government adoption: architecture and barriers [J]. Business Process Management Journal, 2005, 11(5): 589-611.

[19] Halachmi A. E-government theory and practice: The evidence from Tennessee (USA) [J]. Frontiers of Public Administration, 2004, 24.

[20] Griffin D, Halpin E. An exploratory evaluation of UK local e-government from an accountability perspective [J]. The Electronic Journal of e-Government, 2005, 3(1): 13-28.

[21] Kavanaugh A L, Isenhour P L, Cooper M, et al. Information technology in support of public deliberation [M]. Communities and Technologies 2005. Springer. 2005: 19-40.

[22] Veljković N, Bogdanović-Dinić S, Stoimenov L. Benchmarking open government: An open data perspective [J]. Government Information Quarterly, 2014, 31(2): 278-90.

[23] Kroski E. On the move with the mobile web: libraries and mobile technologies [J]. Library technology reports, 2008, 44(5): 1-48.

[24] Wagner C, Cheung K, Lee F, et al. Enhancing e-government in developing countries: managing knowledge through virtual communities [J]. The Electronic Journal of Information Systems in Developing Countries, 2003, 14(

[25] Aurigi A. Making the digital city: the early shaping of urban internet space [M]. Ashgate Publishing, Ltd., 2005.

[26] Mutula S M. Assessment of the e-readiness of small and medium sized enterprises in the ICT sector in Botswana, with special reference to information access [D]; University of Johannesburg, 2005.

[27] Davin E, Majidi N. Study on cross border population movements between Afghanistan and Pakistan [J]. Commissioned by the Office of the United Nations High Commissioner for Refugees (UNCHR), Kabul, 2009,

[28] Bertot J C, Jaeger P T, Hansen D. The impact of polices on government social media usage: Issues, challenges, and recommendations [J]. Government Information Quarterly, 2012, 29(1): 30-40.

[29] Alsufayri K. Universalising electronic government services: facing the digital divide challenge [D]; Auckland University of Technology, 2014.

[30] Sone J W. E-Governance in central Texas: patterns of e-gov adoption in smaller cities [D]; Texas State University, 2011.

[31] Anna Xiong J. Current status and needs of Chinese e-government users [J]. The Electronic Library, 2006, 24(6): 747-62.

[32] Bonsón E, Torres L, Royo S, et al. Local e-government 2.0: Social media and corporate transparency in municipalities [J]. Government information quarterly, 2012, 29(2): 123-32.

[33] Zambrano R. E-governance and development: Service delivery to empower the poor [J]. Social and Organizational Developments through Emerging E-Government Applications: New Principles and Concepts: New Principles and Concepts, 2009, 98.

[34] Ifinedo P. Factors influencing e-government maturity in transition economies and developing countries: a longitudinal perspective [J]. ACM SIGMIS Database, 2012, 42(4): 98-116.

[35] Carmichael L R, Mcclure C R, Mandel L H, et al. Broadband Adoption| Practical Approaches and Proposed Strategies for Measuring Selected

Aspects of Community-Based Broadband Deployment and Use [J]. International Journal of Communication, 2012, 6(22.

[36] Owens C. Communicating an organisation's identity to library users: a case study within the New Zealand community library sector [D]; Unitec Institute of Technology, 2013.

[37] Susha I. Participation in open government [J]. 2015,

[38] Walts N. Native American Indian tribal college and university students: A qualitative study of the digital divide [D]; UNIVERSITY OF PHOENIX, 2011.

AUTHOR PROFILE

Zulfiqar Haider (Syed Zaidi) is working as an Assistant Professor in the Department of Public Administration, Faculty of Social Sciences, University of Sindh, Jamshoro, Pakistan. He is currently pursuing his Ph.D. degree at Dalian University of Technology, Dalian, China.

Prof. Chen Shuwen is working as a Dean at the School of Public Administration and also performing his duties as a doctoral programme supervisor at Dalian University of Technology, Dalian, China.

Prof. Dr. Farah Lalani is working as a Professor in the Department of Public Administration, Faculty of Social Sciences, University of Sindh, Jamshoro, Pakistan.

Dr. Aftab Ahmed Mangi is working as an Assistant Professor in the Department of Public Administration, Faculty of Social Sciences, University of Sindh, Jamshoro, Pakistan.

Building a Robust Client-Side Protection Against Cross Site Request Forgery

Abdalla AlAmeen
College of Art and Science- Prince Sattam bin Abdulaziz University

***Abstract*—In recent years, the web has been an indispensable part of business all over the world and web browsers have become the backbones of today's systems and applications. Unfortunately, the number of web application attacks has increased a great deal, so the matter of concern is securing web applications. One of the most serious cyber-attacks has been by cross site request forgery (CSRF). CSRF has been recognized among the major threats to web applications and among the top ten worst vulnerabilities for web applications. In a CSRF attack, an attacker takes liberty be authorized to take a sensitive action on a target website on behalf of a user without his knowledge. This paper, providing an overview about CSRF attack, describes the various possible attacks, the developed solutions, and the risks in the current preventive techniques. This paper comes up with a highly perfect protection mechanism against reflected CSRF called RCSR. RCSR is a tool gives computer users with full control on the attack. RCSR tool relies on specifying HTTP request source, whether it comes from different tab or from the same one of a valid user, it observes and intercepts every request that is passed through the user's browser and extracts session information, post the extracted information to the Server, then the server create a token for user's session. We checked the working of RCSR extension, our evaluation results show that it is working well and it successfully protects web applications against reflected CSRF.**

Keywords—Security; Reflected CSRF; client-side protection; tab ID; token

I. Introduction

In recent years, the web has been an indispensable part of business all over the world and web browsers have become the backbones of today's systems and applications. Unfortunately, the number of cyber-attacks has increased a great deal, so the matter of concern is securing web applications. One of the most serious attacks has been called cross site request forgery (CSRF). CSRF is also known as XSRF, Session Riding, One-Click-Attack, and Confused Deputy [3]. In a CSRF attack, an attacker takes liberty be authorized to perform a sensitive action on a target website on behalf of a user without his knowledge.

CSRF attacker takes advantages of implicit authentication mechanisms of HTTP protocol and cached credentials in the browser to inject web applications with malicious script [17]. The malicious script may destroy the privacy of the user's session with a web application. CSRF attack tricks user's browser into performing requests into a target web site that is vulnerable to CSRF [4]. A website is vulnerable to CSRF attack when it has inadequate mechanism to check whether a valid request has been sent intentionally or unintentionally by a logged in user [15]. A CSRF attack involves three actors as shown in Fig 1, a user, a trusted website, and a malicious website. To perform a CSRF attack, the user must hold an active session with the target site [13]. Suppose the victim user is authenticated (a logged in user), the attacker can upload HTML element or JavaScript code on a third-party website, subsequently the victim user visits an attacker controlled third-party website or he/she clicks on a link in the same web browser (without logging out form the trusted website). Thus, the attacker malicious script will be executed without the victim user being aware of it. Attacker uses illegal strategies to deceive the victim to send unintended request [2]. For instance, an attacker may attract browser's user into clicking on a malicious link or image, which is hosted on untrusted third party server or he/she can post a message in a social website, this message may contain malicious image tag as shown in Listing 1.

```
<img src="http://mybank.com/withdraw?
account=Sender&amount= amount-&for= reciever ">
```

Listing 1. Image tag containing a malicious Code snippet

As shown in Listing 1, the attacker may send an image tag a third-party website, that contains a request to perform a sensitive action (withdraw money) on a trusted-website of an authenticated user (mybank.com), probably without their knowledge.

In the early appearance of World Wide Web (WWW) in 1989 [12], it only contains a set of static pages interconnected via hyperlinks. But when images were added to web pages in 1993 [12], a request to a web page could cascade a set of requests to multiple other web pages. Thus, cross-site or cross-origin requests triggered without explicit user interaction. With the coming of interactive web thought Java scripts and Web forms in 1995 [10], cross-site interactions become a real security threat to web applications.

Typically, today's websites implement cookies to identify authenticated users [1]. After the user is successfully authenticated by the Web server, the browser will get an identity login cookie to remember the logged-in status [10]. Later, when the user is visiting the Web pages of the target website, the browser will automatically attach the identity login cookie in the HTTP request [10]. This cookie will not be removed until the browser is closed or the user is logged out. The attacker is able to abuse this duration to make some user's browser perform authenticated requests probably without their knowledge, and that is what is called cross site request forgery CSRF.

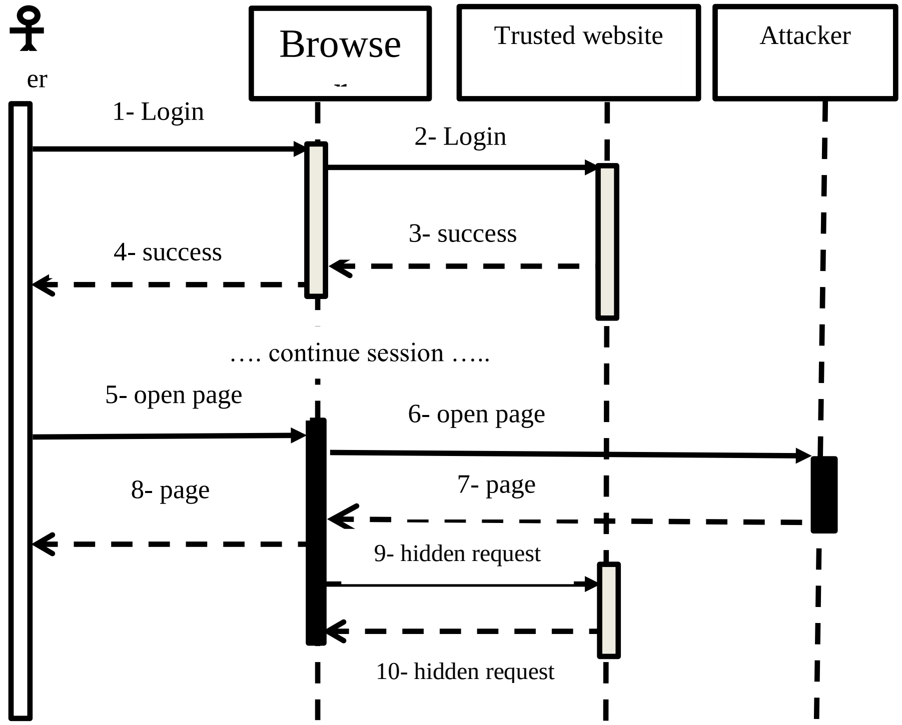


Fig. 1. Simple CSRF attack scenario

A number of serious CSRF vulnerabilities in some websites were documented [7], which allowed an attacker to transfer funds from victim's account to an account chosen by the attacker [7] as shown in listing 1. What makes detection or prevention of CSRF attack so difficult is the fact that web applications look to all requests triggered from an authenticated user's browser just like another. Since the requests are being made directly to the real Web application (no man-in-the middle) therefore the unintended malicious requests are considered legitimate in server perspective: “The only problem is the victim did not intend to make the request, but the Web server does not know that” [12]. The majority of web application are vulnerable with users having very little ability to defend themselves against CSRF” [12].

One of the primary causes of CSRF attacks is the misuse of cached credentials in cross-domain requests [7]. The attacker can easily send some requests to web applications in another trusted web site without the user involvement and knowledge. This makes web browser send cross-site requests, while implicitly using cached credentials in web browser [7].

CSRF attacks are as powerful as a user. Whatever action that the user can do can also be done by an attacker using a CSRF attack. Thus, the more rights a site gives to a user, the more dangerous are the possible CSRF attacks. The seriousness of CSRF attack comes from the fact of malicious request arriving from authenticated user. For instance, if the account of the target has full rights, this can destroy the overall web application. However, if we can understand all the steps in which Web applications are attacked via CSRF attacks, we can design countermeasures to thwart it. Moreover, if we know who the attackers are, and what they want, their goals, motivations and abilities we will have to educate users to protect themselves from CSRF attacks.

The main aim of this paper is to follow preventive techniques in order to make web application more secure than it is at present .This paper, however, provides an overview about CSRF attack, the various possible attacks, the developed solutions, and the risks in the current preventive techniques. This paper comes up with a highly perfect protection mechanism against reflected CSRF. RCSR is a tool that gives computer users full control on the attack. RCSR tool relies on specifying HTTP request source whether coming from different tab or from the same one of a valid user. RCSR observes and intercepts every request that is passes through the user’s browser. RCSR extracts the session information such as tab ID, IP address, then post the extracted information to the web server, the server creates a token for user's session to validate the legitimacy of the request before changing any sensitive data in the server database. We have checked the functioning of RCSR extension, our evaluation results shows that it is working well and it successfully protects web applications against reflected CSRF.

The remainder of this paper is structured as follows: Section 2 describes the main concepts of CSRF and the processes involved in the attacks. Section 3 describes the existing protection and prevention techniques against CSRF. Section 4 focuses entirely on the development of tokens concepts as a standard defence mechanism against CSRF. Section 5 summarizes some existing defence's techniques and their attributes. Section 6 presents RCSR, our proposed scheme, section 7 describes the implementation of RCSR. In section 8, we extensively validate the efficiency and the capability of the RCSR tool against reflected CSRF attack and finally in Section 9 we conclude this paper.

II. Cross Site Request Forgery

The launching of CSRF attack may be carried out in different steps depending on the type of CSRF attack. CSRF can mainly be classified into two types: reflected and stored [14]. First of all, the attacker should know the structure of the website request forms, then check the main functionality of targeted web site. A professional attacker may perform that manually or by searching the web using specific software tools. Toolkits such as seobook, webconfs and web spider are the software available on the web for free .They can be used for displaying the contents of a web-page and its functionality. Secondly, the attacker will specify specific functionality in the web-page that it can be used to perform malicious actions on behalf of a victim user. Then, the attacker will send a parameterized request. Some network protocol analyser such as Wireshark, Cain & Abel and Tcpdump can be used to examine data from a live network and browse the captured data that may contain buttons or links that can perform actions. The following step is to create a malicious link that can send this legitimate HTTP request to the website and will execute some interesting functionality on the server such as transferring money, changing a password, etc. Finally, the attacker needs to convince a logged in user into the target website to click on the malicious link to execute the CSRF attack successfully.

For launching reflected CSRF attacks, the attacker needs to include the malicious link on the attacker's controlled website and trick the user to click on the link, or where an XMLHTTPRequest object may automatically execute the attack when a user visits the website [14].

For stored CSRF attacks, the attacker needs to create some posts that embed the malicious link in the target website, or execute a stored XSS attack on a website where an XMLHTTPRequest object will automatically execute the attack as soon as a user visits the page [14]. This removes the step of convincing a user to click on a link.

III. Existing Countermeasures

To overcome CSRF attacks, a variety of techniques are available to protect server applications and the end-users from CSRF attack [7]. CSRF protection and prevention techniques can be classified into two main categories:

1) *Client side protection techniques*
2) *Server side protection techniques*

Client side protection techniques can be used to protect users from CSRF attacks by monitoring outgoing requests and incoming responses. Client side protection techniques can be implemented as a browser proxy (plug-in or extension) to web browsers [19]. Browser extension is the technique that we have adopted in this paper as shown in section (6).

The basic idea behind the Server side protection techniques is that server can strip authentication credential and session information from suspected requests, or it can refuse such requests. Using validation of secret token and checking HTTP Referrer header are the most applied Server side protection techniques [7]. Unfortunately, not any of the proposed mechanisms is fully capable of carrying out this task, in other words the existing solutions are time-consuming, error-prone, and not immune to avoid CSRF attacks.

IV. CSRF Tokens Concept

In the early appearance of World Wide Web in 1989 [12], it only contained a set of static pages interconnected via hyperlinks. But when images were added to web pages in 1993 [12], a request to a web page could cascade a set of requests for other multiple web pages. Thus, cross-site or cross- origin requests triggered without explicit user interaction. With the coming of interactive web thought Java scripts and Web forms in 1995 [10], cross-site interactions has become a real security threat to web applications.

Typically, today's websites implement cookies to identify authenticated users [1]. After the user is successfully authenticated by the Web server, the browser will get an identity login cookie to remember the logged-in status [a10]. Later, when the user is visiting the web pages of the target website, the browser will automatically put the identity login cookies in the request [a10]. This will not be removed until the browser is closed or the user logged out.

CSRF vulnerabilities arise because the browsers send the cookies back to the Web server automatically with each subsequent request. If Web applications relied solely on cookies as a mechanism to keep track of user sessions, they will be at risk for this type of attack [12].

The attacker is able to abuse this duration to make the user's browser perform authenticated requests probably without their knowledge, and that causes what is called CSRF.

To create a standard defence mechanism against CSRF attack, we must support cookie-based and HTTP authentication with additional means to keep track of sessions. These additional means may be as additional tokens that are transmitted through hidden fields with any request to the server [12].

When the server receives a request, in addition to the process of verifying the validity of session cookies, it also verifies that the received token is valid for the current user, else the request will be rejected. If we assume that the attacker does not have the ability to know the value of this token, so he won't be able to put the right token in its submissions, therefore he does not have the ability to launch a successful CSRF attack. When we use CSRF tokens in this way, they must be subject to adequate protection because they are considered as sensitive data [12]. If the attacker can predict the value of CSRF tokens that have been sent to another user, so he can obtain a valuable data to perform malicious action on behalf of the user.

V. Literature Review

To overcome CSRF attacks, a variety of defense techniques exist, these protections and prevention schemes propose to make forgery requests harder for adversaries, or to confirm the origin of page requests. By assuring the integrity of requests' origins, defense techniques can ignore page requests coming from the cross site domains because web transactions are usually intended for the requests initiated from the same domain, not the cross sites. Most defense schemes can be classified into two main types, Client side techniques and Server side techniques. Client-side defense protects users even if web application doesn't prompt in fixing their vulnerabilities. Moreover, they have accurate information about the requests sources, whether they result from clicking on a link, or a bookmark on a trusted web page. There are several Client- side protection and prevention schemes (Secret Validation Token, RequestRodeo, CsFire etc.) to prevent CSRF attack.

A. *Secret validation token*

Secret Validation Token is a well-known client- side protection scheme against CSRF attacks. Not like other verification, token approach does not require user intervention, so the users will not know that something has been used to protect them. This scheme sends additional information with each HTTP request to determine whether the request came from an authorized user. To apply token, web applications must first create a “presession,”, and then proceed forward a real session after successful authentication [18]. Validation token should be hard to guess for attacker who does not already have access to the user’s account. If a request is missing a token or the token does not match the expected value, the web application should reject the request and prompt the user [18].

One disadvantage of using the Secret Validation Token is that, occasionally, some users disclose the contents of web pages they view to third parties, for instance via email . If the page contains the user’s Validation Token, anyone who views the contents of the page can impersonate the user to the web application until the session expires [19].

B. *Requestrodeo*

Johns and Winter proposed RequestRodeo as a client-side protection proxy against CSRF [11]. RequestRodeo lies, next to cookie-based HTTP authentication. This technique offers protection via detecting cross-domain requests and then removal of cookie values from these requests (stripping of implicit authentication) [11]. A request is authenticated when it satisfies the Same Origin Policy (SOP) and it initiated as a result of an interaction with the currently viewed browser' tab. RequestRodeo is limited to only certain HTTP requests and no HTTPS requests, so it does not scale well to web 2.0 applications. RequestRodeo fails to detect all JavaScript dynamic links in the responses, since this dynamic content has come after passing through the proxy. Also, RequestRodeo does not differentiate between malicious and genuine cross origin requests, so it provides very poor protection against CSRF [16].

C. *CSFIRE*

CsFire [8] is integrated extension into Mozilla browser to mitigate CSRF attacks, it extends the work of Maes et al. [8], CsFire is the only system that provides formal validation through bounded model checking to defend against CSRF in the formal model of the web developed by Akhawe et al. [8]. CsFire strips cookies and HTTP authorization headers from a cross-origin request. The advantage of stripping cookies and HTTP authorization headers is that there are no side-effects for cross-origin requests that do not require credentials in the first place.

Additionally, CsFire supports users for creating custom policy rules, which use user-supplied whitelist and blacklist to certain traffic patterns. Furthermore, CSFire utilizes a sophisticated heuristic to identify legitimate cross-domain requests which are allowed to carry authentication credentials [8]. The disadvantage of CsFire approach is that without the server supplied or user supplied whitelist, it will not be able to handle complex, genuine cross origin scenarios and the whitelists need to be updated frequently.

According to the defensive techniques discussed above, none of them is able to provide full protection. So it seems necessary to overcome the drawbacks of present defensive measures. We propose to develop a new client side defensive approach, in the form of a Firefox extension, to prevent Reflected CSRF attacks effectively as explained in section 6.

VI. THE PROPOSED SCHEME

The web browser is the right place to apply appropriate protection mechanism for web application because it is the first place to detect CSRF attack symptoms. So the proposed defense mechanism against reflected CSRF attacks should be applied on the client side in order to reduce the overtime efforts of web developers. Client side protection techniques can be implemented as a proxy or as plug-in (extension) to web browsers.

Mozilla is an extensible architecture, open source, and the second most popular browser, also Mozilla browser behaviour can be modified by creating appropriate XPCOM (Cross Platform Component Object Model) objects and implementing a set of APIs. Mozilla supports the global browser object called (gBrowser) to access the active tab windows and examine its ID through GetSelectedTab function. We propose to provide a robust client side defense mechanism against CSRF, hence named as "Robust Client Side Request‟ (RCSR). Once implemented on the browser, RCSR can be the best solution over other techniques to protect web applications against reflected CSRF. RCSR is a technology independent tool and does not depend on user input, so it solves the drawbacks of current protection techniques.

We designed the plug-in using JavaScript, which can be installed in Mozilla browser to protect users against reflected CSRF attacks. A user needs to enable CSRF from the tools menu of a browser after loading a page that needs to be monitored for attack detection.

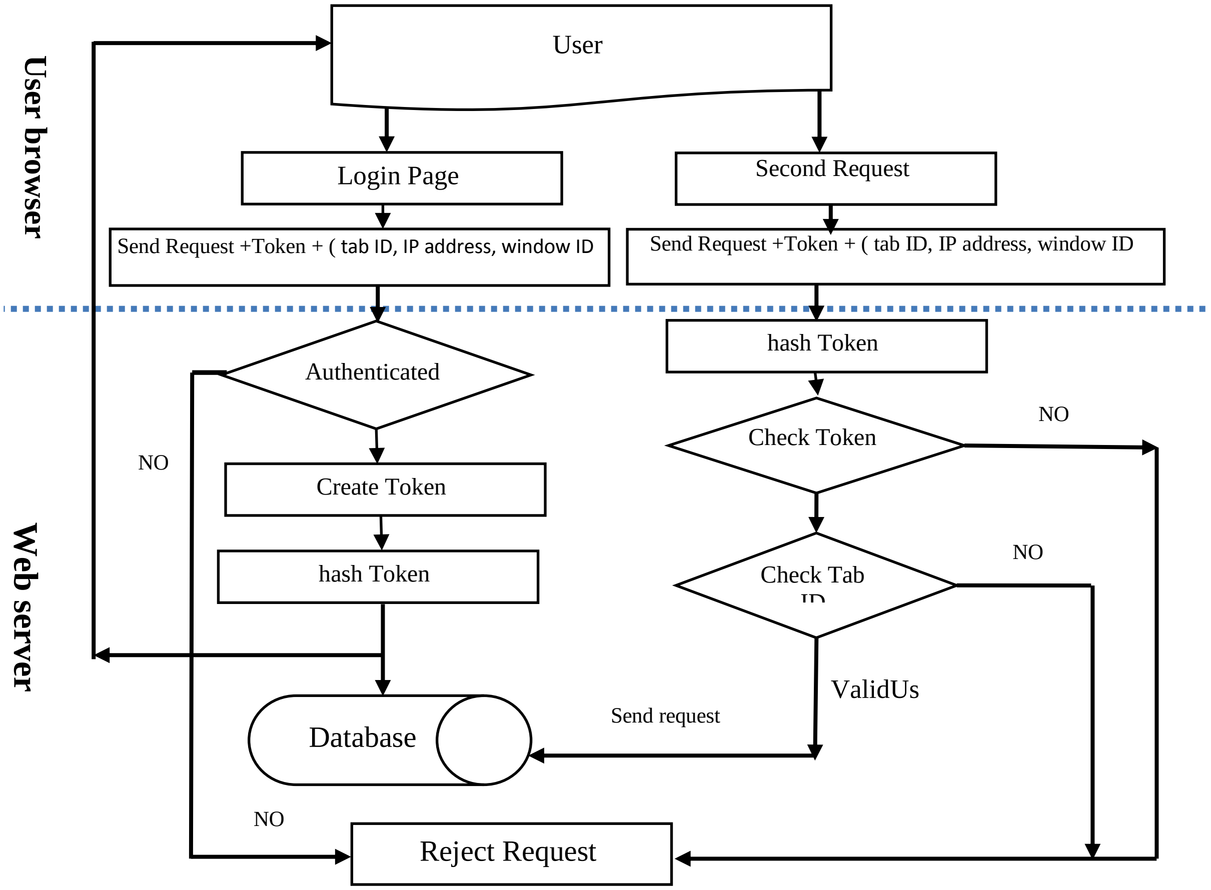


Fig. 2. The proposed scheme diagram

VII. Implementation

Our solution, RCSR tool a simple policy, is implemented in the form of a client-side plug-in to protect web applications against reflected CSRF. In general, RCSR allow the web application developer to plug in new functionalities to web browser.

The general mechanism of RCSR functions as follows:

To specify HTTP request source, whether coming from different tab or from the same one of a valid user, RCSR observes every request that is passes through the user's browser, intercepts HTTP requests and extracts session information. Listing 1 below show snippet code to extract the tab ID from web browser.

```
function RCSRObserver()
                  {
  this.register();
  this.windowIds = new Object();
                  }
RCSRObserver.prototype =
                  {
        observe : function(subject, topic, data)
                  {
   var tabId = this.getTabIDfromDOM(httpChannel, subject);
                 if (tabId)
                  {
        var windowId = this.windowIds[tabId];
                 if ( ! windowId)
                  {
              this.registerTab(tabId);
           windowId = this.windowIds[tabId];
                  }
 httpChannel.setRequestHeader("Window-Id", windowId,
                  false);
                  }
                  }
   getTabIDfromDOM : function(aChannel, aSubject)
                  {
           registerTab : function(tabId)
                  {
         window.addEventListener("load",
                  function(e)
                  {
         observer = new RCSRObserver();
      var num = gBrowser.browsers.length;
         for (var i = 0; i < num; i ++ )
                  {
      var b = gBrowser.getBrowserAtIndex(i);
    observer.registerTab(b.parentNode.id, i);
                  }
```

Listing 2. Extracting tab ID Code snippet

To create a new token for the current user session, RCSR post the extracted information (tab ID, IP address, window ID) to the web Server.

The system will store the session information on the server database to map token with user's session or identity. The web server will send to the client with a unique token. After hashing the information by a cryptographic hash algorithm based on SHA-1 the web server will define the tab ID value. Then store the hashed value in database table. The web application can repeatedly validate the legitimacy of the attached tokens before changing any sensitive data in the server database.

Sever verifies if the request is generated from the same tab of the browser. This verification is performed by comparing the stored hash information with the hashing information that is sent with each request. The request will be executed if the comparison result is true, otherwise the session will be destroyed. Fig. 2 shows all steps of the RCSR detection tool.

VIII. Evaluation

We conducted some tests to evaluate the efficiency and the capability of the RCSR tool against reflected CSRF attack, to make sure that its results match with what is predetermined, to discover the problem and try to fix before the deployment.

PhpBB 3 is an open source discussion forum software. IT includes all the features in today's top of the line software written in PHP, and MySQL [9]. PhpBB 3 uses cookies to authenticate user's requests which are an important element in CSRF attack. Despite the fact that phpBB3 is popular application and well-maintained, but easily we discovered some CSRF vulnerabilities [9].

By exploiting CSRF attacks, we modified some important information through abusing of an authenticated user privileges. Through malicious link we could access the user cookies and valid session on the victim's browser, so we were able to send and delete some messages from the forum or even change user name and password on behalf of the victim user.

To evaluate the ability of RCSR to protect vulnerable applications, we installed RCSR tool as an extension to Mozilla browser. When we repeated the previous attacks, RCSR tool detected and rejected all CSRF attempt correctly.

While testing, that RCSR doesn't interfere with the normal application behaviour. We observed and compared phpBB3 application behaviour without the RCSR tool protection to the behaviour with enabled CSRF protection. The results were identical and the RCSR tool succeeded in performing its task transparently.

We tested some functionalities of the Mozilla browser after installing RCSR tool. For instance, we observed the correct behaviour of the Mozilla's “Back” and "Forward" button, which is a widely used convenience feature that must not be broken by CSRF protection.

To test RCSR performance, we observed no noticeable delay when interacting with the applications protected by RCSR.

IX. Conclusion

One of the most serious cyber-attacks has been by cross site request forgery (CSRF). CSRF has been recognized among the major threats to web applications and among the top ten worst vulnerabilities for web applications. In a CSRF attack, an attacker takes liberty be authorized to take a sensitive action on a target website on behalf of a user without his knowledge. To conclude this paper, we have discussed CSRF in different domains, the severity of the attack on the current web applications, the various possible CSRF attack and risks in the current preventive techniques. To overcome the drawbacks of present defensive protection, this paper proposed a new client side defensive tool (RCSR). RCSR is a Firefox extension, which can prevent Reflected CSRF attacks effectively. RCSR is a tool gives computer users with full control on the attack. RCSR tool relies on specifying HTTP request source, whether it comes from different tab or from the same one of a valid user, it observes and intercepts every request that is passed through the user's browser and extracts session information, post the extracted information to the Server, then the server create a token for user's session.

In a practical evaluation, the working of this extension was checked against reflected CSRF, the evaluation results show that it is working well. It successfully protects web applications against reflected CSRF. In future work we plan to extend the RCSR functionality against stored CSRF attacks and evaluate its performance to make it more powerful and accurate. Finally, we hope that RCSR tool will prove useful in protecting web applications.

Acknowledgment

This project was supported by the Deanship of Scientific Research at Prince Sattam Bin AbdulAziz University under the research project 2014/01/1660

References

[1] Barth, Adam, Collin Jackson, and John C. Mitchell. "Robust defenses for cross-site request forgery." In Proceedings of the 15th ACM conference on Computer and communications security, ACM, 2008, pp. 75-88.

[2] Batarfi, Omar A., Aisha M. Alshiky, Alaa A. Almarzuki, and Nora A. Farraj. "CSRFDtool: Automated Detection and Prevention of a Reflected Cross-Site Request Forgery." (2014).

[3] Calafato, Andrew, and Kostantinos Markantonakis. "An analysis of the vulnerabilities introduced with the java card 3 connected edition." PhD diss., Msc thesis, Royal Holloway, University of London, 2012.

[4] Chen, Eric Y., Sergey Gorbaty, Astha Singhal, and Collin Jackson. "Self-exfiltration: The dangers of browser-enforced information flow control." InProceedings of the Workshop of Web, vol. 2. 2012.

[5] Chin, Erika, Adrienne Porter Felt, Kate Greenwood, and David Wagner. "Analyzing inter-application communication in Android." In Proceedings of the 9th international conference on Mobile systems, applications, and services, ACM, 2011, pp. 239-252.

[6] Czeskis, Alexei, Michael Dietz, Tadayoshi Kohno, Dan Wallach, and Dirk Balfanz. "Strengthening user authentication through opportunistic cryptographic identity assertions." In Proceedings of the 2012 ACM conference on Computer and communications security, ACM, 2012, pp. 404-414.

[7] De Ryck, Philippe, Lieven Desmet, Thomas Heyman, Frank Piessens, and Wouter Joosen. "CsFire: Transparent client-side mitigation of malicious cross-domain requests." In Engineering Secure Software and Systems, Springer Berlin Heidelberg, 2010, pp. 18-34.

[8] De Ryck, Philippe, Lieven Desmet, Wouter Joosen, and Frank Piessens. "Automatic and precise client-side protection against CSRF attacks." InComputer Security–ESORICS 2011, Springer Berlin Heidelberg, 2011, pp. 100-116.

[9] Fong, Matthew, Herman Lee, Chih-Hao Lin, and David Yue. "Security Analysis of phpBB3 Bulletin Board Software." (2010).

[10] Hall, Marty, and Larry Brown. Core Web Programming. Prentice Hall PTR, 2001.

[11] Johns, Martin, and Justus Winter. "RequestRodeo: Client side protection against session riding." In Proceedings of the OWASP Europe 2006 Conference. 2006.

[12] Kappel, Gerti, Birgit Pröll, Siegfried Reich, and Werner Retschitzegger. Web engineering. John Wiley & Sons, 2006.

[13] Poyar, Ryan L. "Cross-site request forgery attacks against Linksys wireless routers." (2010).

[14] Shahriar, Hossain, and Mohammad Zulkernine. "Client-side detection of cross-site request forgery attacks." In Software Reliability Engineering (ISSRE), 2010 IEEE 21st International Symposium on, IEEE, 2010, pp. 358-367.

[15] Singh, Nanhay, Achin Jain, Ram Shringar Raw, and Rahul Raman. "Detection of Web-Based Attacks by Analyzing Web Server Log Files." In Intelligent Computing, Networking, and Informatics, Springer India, 2014, pp. 101-109.

[16] Telikicherla, Krishna Chaitanya, Venkatesh Choppella, and Bruhadeshwar Bezawada. "CORP: A Browser Policy to Mitigate Web Infiltration Attacks." InInformation Systems Security, Springer International Publishing, 2014, pp. 277-297.

[17] Wedman, Shellie, Annette Tetmeyer, and Hossein Saiedian. "An analytical study of web application session management mechanisms and HTTP session hijacking attacks." Information Security Journal: A Global Perspective 22, no. 2 (2013), 55-67.

[18] Xing, Xinyu, Elhadi Shakshuki, Darcy Benoit, and Tarek Sheltami. "Security analysis and authentication improvement for ieee 802.11 i specification." InGlobal Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE, 2008, IEEE, pp. 1-5.

[19] Zeller, William, and Edward W. Felten. "Cross-site request forgeries: Exploitation and prevention." The New York Times (2008): 1-13.

Content -based Image Retrieval for Image Indexing

Md. Al-Amin Bhuiyan
Department of Computer Engineering
King Faisal University
Hofuf, Al-Ahsa 31982, Saudi Arabia

***Abstract*—Content-based image retrieval has attained a position of overwhelming dominance in computer vision with the advent of digital cameras and explosion of images in the Internet and Clouds. Finding the most relevant images in a short time is a challenging job with many big cloud sites competing in image search in terms of accuracy and recall. This paper addresses an image retrieval system employing color information indexing. The system is organized with the hue components of the HSV color model. To assess the precision of the image retrieval system, experiments have been carried out on a database consisting of 450 images drawn by Japanese traditional painters, namely Sharaku, Hokusai, Hiroshige, and the images obtained from the World Wide Web (WWW) multicolor natural scenes. In order to query the database, the user specifies an object on which the same color attributes are evaluated and all similar looking images are exposed as the outcomes of the query.**

***Keywords*—color indexing; HSV color model; color histogram; Minkowski distance metric; fuzzy clustering; Color Quantization**

I. INTRODUCTION

Content-based image retrieval is emerging tremendous interest in image processing, computer graphics, computer vision, pattern recognition, image management system, and so on [1-4]. In distinction to the traditional text-based approach, several benefits have been reported in literature for content-based access to images, such as automatic identification, classification, recognition, and retrieval of large digital libraries with photographic images, satellite images, medical images for remote searching and browsing over the ever increasing World Wide Web (WWW).

A fair amount of developments were carried on over the last couple of decades in image retrieval system due to the enormous interest of establishing multimedia information systems and database systems. The convergence of image processing/computer graphics and database technology yields the basis for the creation of such digital image archives. Moreover, with the proliferation of WWW, a remarkable amount of visual information is ready accessible publicly. As a result, it has become a promising demand for search strategies retrieving pictorial entities from large image documentations [5]. Over the past half century there has emerged an increasing interest in the cultural heritage of Japanese society. In tandem with this, and indeed a logical consequence of this, Japanese traditional painting pictures, known as Ukiyoe pictures [6], have been a growing concern with preserving, disseminating, displaying and effectively exploiting the rich cultural resources embodied in many museum and art gallery collections. Benefits of the technology are numerous and the most important points which are generally given are that the use of digital versions of surrogate representations of works of art can: assist security, provide a central database of information to provide easy retrieval of relevant material, assist in preservation of originals and provide a networkable resource of images which greatly increases availability and access to the system.

Image contents include color, shape and texture. Among these contents of images, color provides an efficient visual clue for image retrieval. Managing image data in this regard entails processing, storage, and retrieval of pictorial representations [5]. Due to its graphical illustration, the color histogram becomes the most frequently used technique for image indexing. It provides a convenient tool for computing the similarity between different images, since it proves to be robust to object translation, scaling, rotation, occlusion, deformation, and so on [7].

A substantial amount of research works have been reported in literature [8-16] on content-based image retrieval (CBIR). Swain and Ballard [17] proposed a color-based object recognition employing color histograms for matching between image regions and query objects. Kieldson and Kender [18] applied Gaussian kernels to smooth the histograms on finding skin in color images. Funt and Finlayson [19] developed a color indexing algorithm for object recognition to take into account the influence of lighting conditions. Ennesser and Medioni [20] proposed a local histogram method to localize objects in images. Chang and Wang [21] developed a texture segmentation algorithm employing color histogram. McKenna, Raja and Gong [22] employed an adaptive Gaussian mixtures to model the color allocations of objects. Androutsos, Plataniotis and Venetsanopoulos [23] established a cosine metric based distance measure for color image indexing and retrieval. Their query method is very flexible and provides single and multiple color queries. Liu and Ozawa [24] proposed the spatial neighborhood adjacency graph (SNAG) which could serve as a basis for detecting object by color contents from the candidate images. Sharma et al. [25] have represented images by global descriptor and developed a CBIR system that used color histogram processing. This system is not yet a commercial success. Because the distribution of RGB values changes proportionally with the illumination, which is suitable to some images but have low precision on others.

This paper addresses a color-histogram based method for indexing and retrieving color images. Different dominant and perceptually relevant colors have been extracted from each image in RGB model and are stored in the respective database files. Images are being identified and classified in HSV space depending on the color contents prevailing in these dominant colors, that is, whether a particular color component is significantly present in an image or not. Similarity between different images has been calculated on the basis of Minkowski distance metric. Experimental results demonstrate that the method is capable of indexing, classifying, and retrieval of images with distinct color properties.

The rest of the paper is organized as follows. Section II describes color model. Section III illustrates histogram and image retrieval. Section IV and Section V describes color quantization and image query, respectively. Section VI highlights experimental results and finally, Section VII draws the overall conclusions of this paper.

II. Color Model

Numerous color models have been justified for color specification, such as CIE (R,G,B), (X,Y,Z), (L*,u*,v*), and so on. The main drawback of the CIE (R,G,B) model is that it is not perceptually uniform and the proximity of colors in RGB space does not indicate color similarity. The (X,Y,Z) color space is not uniform, that is, the Euclidean distance between two colors is not proportional to the color difference perceived by humans. Although the (L*,u*,v*) space is uniform, but nevertheless, is not intimately related to the way in which humans perceive color [21].

A color is represented in HSV color space by the three features: hue, saturation and value. Hue is the characteristic of visual perception that corresponds to color sensation related to the dominant color, saturation indicates the comparative purity of color content and value specifies the brightness of a color. The HSV color model organizes similar colors under similar hue alignments. The transformation from RGB color space to HSV space is given by the equations [23,26-28].

$$H = \cos^{-1}\left\{ \frac{\frac{1}{2}\left[(R-G)+(R-B)\right]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \right\}, \tag{1}$$

ranging [0,2π],

$$S = \frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)}; \tag{2}$$

$$V = \frac{\max(R,G,B)}{255}, \tag{3}$$

where R,G,B are the red, green, and blue component values which exist in the range [0,255].

This research employs HSV color model for classification of pictures drawn by the painters namely Sharaku, Hokusai, Hiroshige and natural pictures. The RGB model has been used to identify the Ukiyoe pictures because these are being distinguished according to their red, green and blue color components in the face parts.

III. Color Histogram and Image Retrieval

Color histogram [29,30] represents the distribution of colors in an image. A color histogram is a stable object illustration which is unaffected by occlusion and changes in viewing conditions, and that a color histogram has the advantage of being insensitive to scaling, rotation, and small deformation of objects and being immune to noise [31].

The basic idea to image retrieval by color content is to extract the characteristic colors from target images which are matched with those of the query. Different images from the database are then searched to check whether a specific color feature value is prominently existing in an image or not. If a number of images contain the substantial amount of that query color, then these are marked according to the priority basis. The architecture of the proposed system is shown in Fig. 1.

Since the hue component of the HSV color model performs better with human chromatic perception [17], the hue component has been chosen to designate the colors of images. Pixels in the image are characterized in the RGB space, so it appears natural to define the color attributes as the red, green, and blue value at each pixel. Let a color image $\mathbf{Q}(x,y)$ consists of three color channels $\mathbf{Q}(x,y) = (Q_R(x,y),$ $Q_G(x,y),$ $Q_B(x,y)),$ or $\mathbf{Q} = (Q_R, Q_G, Q_B),$ at (x,y) of size $M \times N$. A hue histogram $H(i)$ of a color image is achieved by counting the number of pixels which have got a hue value $H(Q_R, Q_G, Q_B) = i$ in the image:

$$H(i) = \frac{\#(H(Q_R, Q_G, Q_B) = i)}{M \times N} \tag{4}$$

where # denotes the number of pixels with a hue value $H(Q_R, C_G, Q_B) = i$ and $M \times N$ is the total number of image locations.

A few sample color images and the hue histograms computed from the respective images are illustrated in Fig. 2.

Images are being classified depending on the prominent colors. Color segmentation has been employed to extract the regions of dominant and perceptually relevant colors. Natural pictures are being separated from those of the painting pictures on the basis of the ratio r_a of the area containing ten dominant colors a_{dc} to that of the total area containing all colors a_{ac}. The reason behind this choice appears from the fact that the painting pictures contain only a few number of colors (the painters use only a limited number of colors during painting) in comparison to that of the infinite number of colors in nature. So the dominant colors contribute more to the images in comparison to those of the natural pictures. Painting pictures are being identified and classified according to the name of the painters, such as Sharaku, Hiroshige and Hokusai depending on the dominant color components because it has been found from the experimental results that the pictures are being fashioned with different colors according to the color choice of the painters. So ten prominent colors are being extracted from the hue histogram in RGB space for each image and the representative vectors are identified as, $\mathbf{Q}_i = (Q_{i,R}, Q_{i,G}, Q_{i,B}), \qquad (i = 1,2,..,N).$ The set

$\mathbf{Q}_1 \cap \mathbf{Q}_2 \ldots \cap \mathbf{Q}_N$ of colors belong to the images that resembles the most widely used colors of a given painter.

The similarity between different painters are calculated on the basis of Minkowski-form vector distance metric from their hue histograms. The generalized Minkowski-form distance metric (L_M norm) is given by:

$$d_M(h_q, h_t) = \left(\sum_{i=1}^{N} \left| h_q^i - h_t^i \right|^M \right)^{\frac{1}{M}} \tag{5}$$

where N is the dimension of the vectors h_q and h_t, and h_q^i is the i-th component of h_q. This research uses $M = 2$, (which is often used for L_M metric).

Let h_q and h_t be the query and target histograms, respectively, then application of the histogram intersection operator introduced in [17] provides a simple way to match two different images I_q and I_t through their color histograms as [32]:

$$H(I_q, I_t) = \sum_{i=1}^{N} \min d_M(h_q, h_t). \tag{6}$$

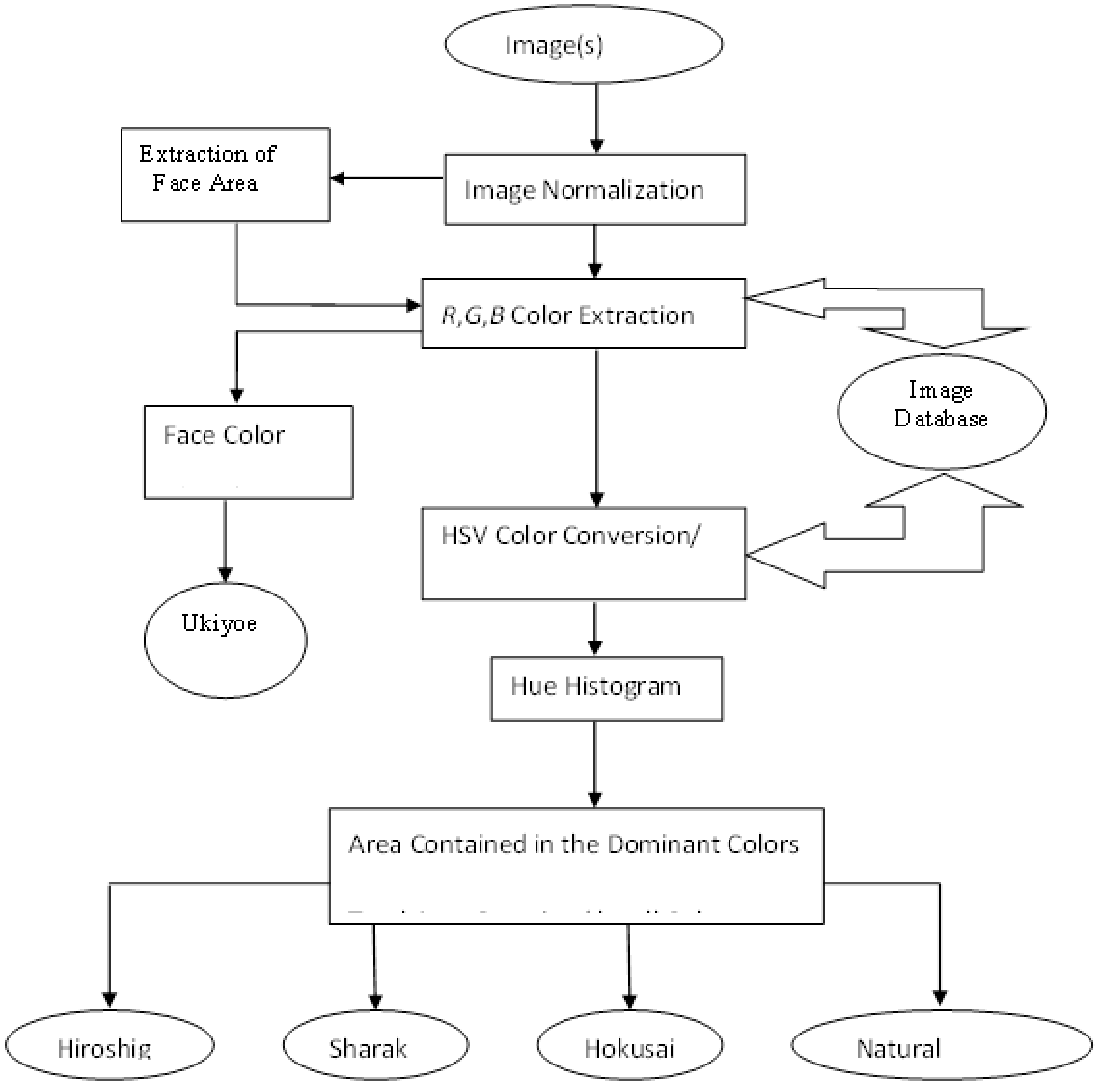
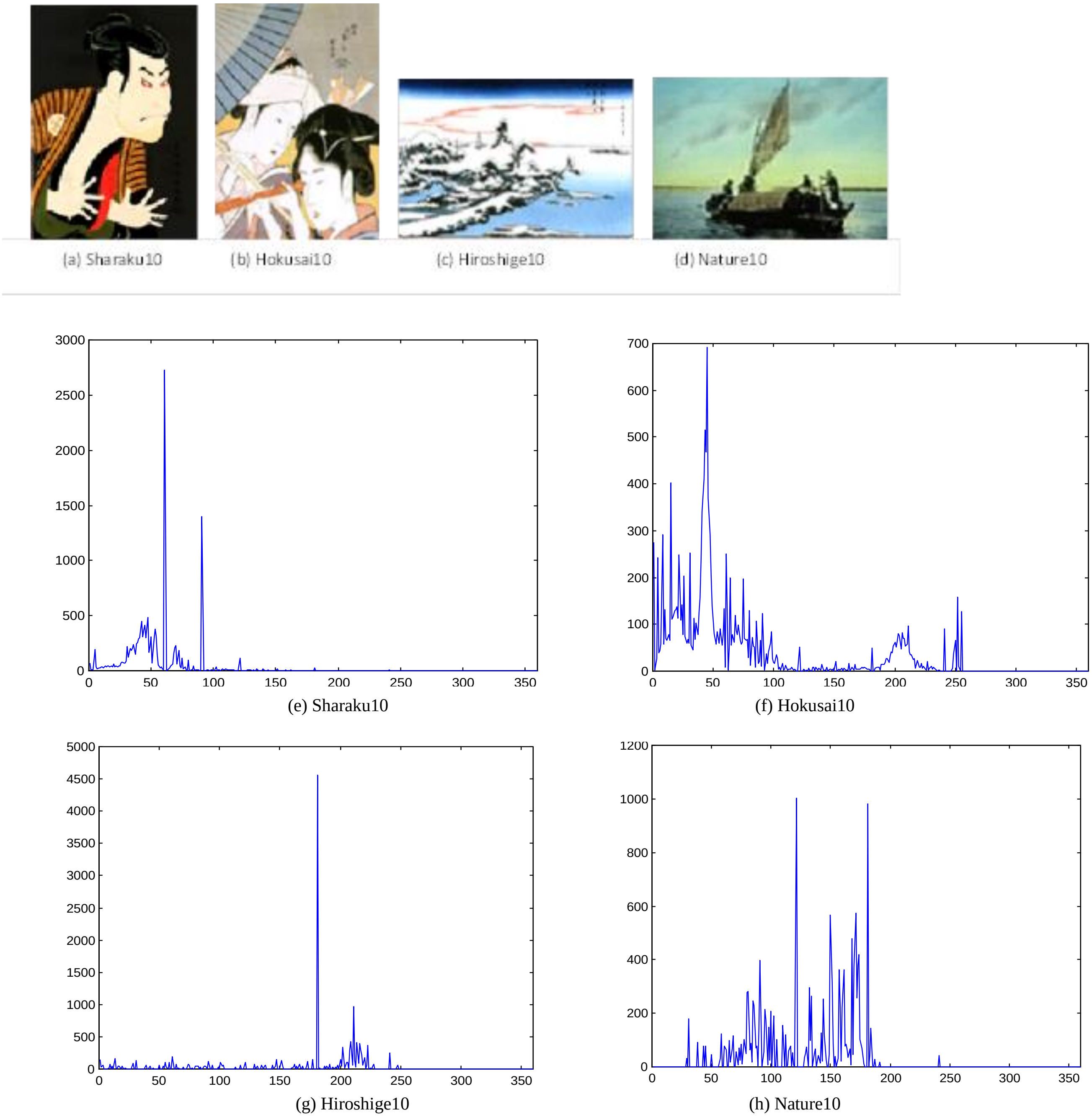


Fig. 1. Architecture of the image retrieval system

Ukiyoe actors are distinguished from those of the natural and painting pictures on the basis of face colors. Human skin colors cluster in a small region in a color space. Although the color representation of a face obtained by a camera is influenced by many factors such as lighting conditions, facial expressions, etc. and the skin colors cluster and differ from person to person in different races [33,34], Ukiyoe actors are nevertheless drawn by some distinct colors by different painters. The presence of some colors in a specific zone provides information whether the images are of actors' faces or not. Fig. 3(a) illustrates a face image, and Fig. 3(b) illustrates the skin color distributions in the RGB color model.

IV. Color Quantization

In order to reduce the computational cost in segmentation, an input color image is quantized so that the number of colors contained in the image is reduced while the primary chromatic information about the image still remains the same. In the quantization method [21], the number of quantized colors are first determined, say *k*, by a histogram thresholding technique. Then fuzzy *c*-means classification algorithm is performed to classify each pixel in the image to one of the *k* colors [34]. The number of clusters are decided depending on the threshold values.


(a) Sharaku10 (b) Hokusai10 (c) Hiroshige10 (d) Nature10

(e) Sharaku10 (f) Hokusai10 (g) Hiroshige10 (h) Nature10

Fig. 2. Hue histogram for different classes of sample images. (a)~(d): original pictures. (e)~(h): the corresponding hue histograms

In fuzzy clustering, each color has a degree of belonging to clusters, rather than belonging completely to just one cluster. Thus, colors on the edge of a cluster, may be in the cluster to a lesser degree than the colors in the center of the cluster. For each color **C** we have a coefficient providing the degree of being in the *j*-th cluster $u_j(\mathbf{C})$. Usually, the sum of those coefficients for any given **C** is defined to be 1:

$$\forall_{\mathbf{C}} \left(\sum_{j=1}^{k} u_j(\mathbf{C}) = 1 \right), \qquad (7)$$

where *k* is the number of clusters.

In fuzzy *c*-means, the centroid of a cluster is the mean of all colors, weighted by their degree of belonging to the cluster. Therefore, the center of the cluster, r_j will be:

$$r_j = \frac{\sum_{\mathbf{C}} u_j(\mathbf{C})^m \mathbf{C}}{\sum_{\mathbf{C}} u_j(\mathbf{C})^m} \qquad (8)$$

where r_j is the center of *j*.

The degree of belonging is related to the inverse of the distance to the cluster center:

$$u_j(\mathbf{C}) = \frac{1}{d(r_j,\mathbf{C})}, \qquad (9)$$

Then the coefficients are normalized and fuzzyfied with a real parameter $m > 1$ so that their sum is 1. Therefore,

$$u_j(\mathbf{C}) = \frac{1}{\sum_i \left(\frac{d(r_j,\mathbf{C})}{d(r_i,\mathbf{C})}\right)^{2/(m-1)}}. \qquad (10)$$

This investigation uses *m* equal to 2, which is equivalent to normalizing the coefficient linearly to make their sum equals 1.

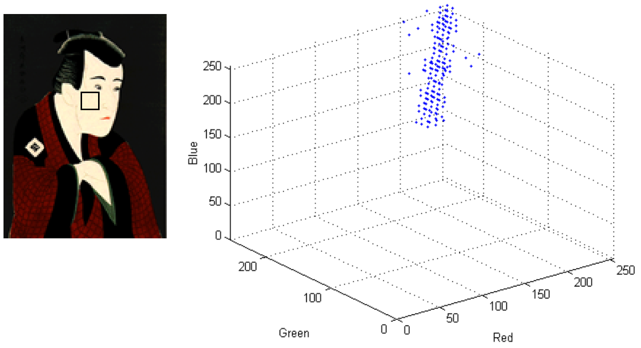


Fig. 3. RGB color distribution of a typical Ukiyoe actor’s face

V. IMAGE QUERY

The query process is to effectively find and retrieve those images from the database that are most similar to the user’s query image. In this case, *z*-nearest neighbor query is employed, which retrieves the *z* images that are most similar to the query image (which are typically sorted by lowest dissimilarity to the query image). Given a number *N* of *I* images and a feature dissimilarity function f_d, find the images $I_T \in N$ such that $f_d(I_q : I_T) \le T_{fd}$, where I_q is the query image and T_{fd} is the threshold for feature dissimilarity. In this case a query returns any number of images depending on the bounds defined by the threshold of feature dissimilarity T_{fd}.

VI. EXPERIMENTAL RESULTS

The effectiveness of the proposed method has been justified over some experimental results. The database furnished for this experiment contains a total of 450 images: 80 drawn by Sharaku, 80 by Hokusai, 80 by Hiroshige and 210 natural pictures (sea, flowers, sunrise, sunset, scenery, animals, architectures, towns, etc.) down-loaded from the Internet. The snapshot of the CBIR software is shown in Fig. 4. When a user selects the query image and specifies the threshold value for L_2 norm, all the similar looking images are then displayed.

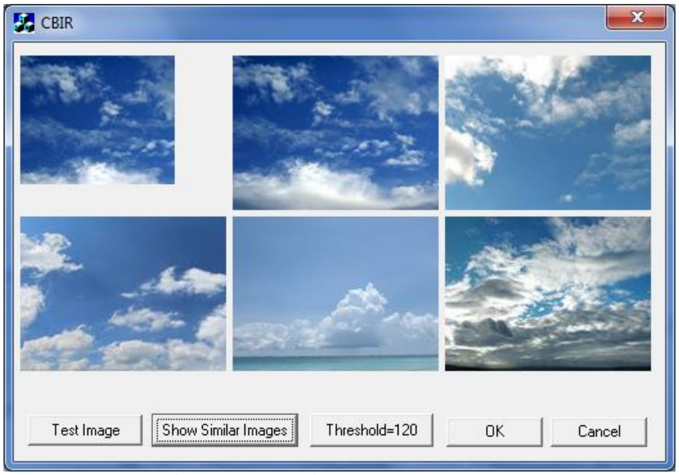


Fig. 4. Snapshot of the CBIR interface

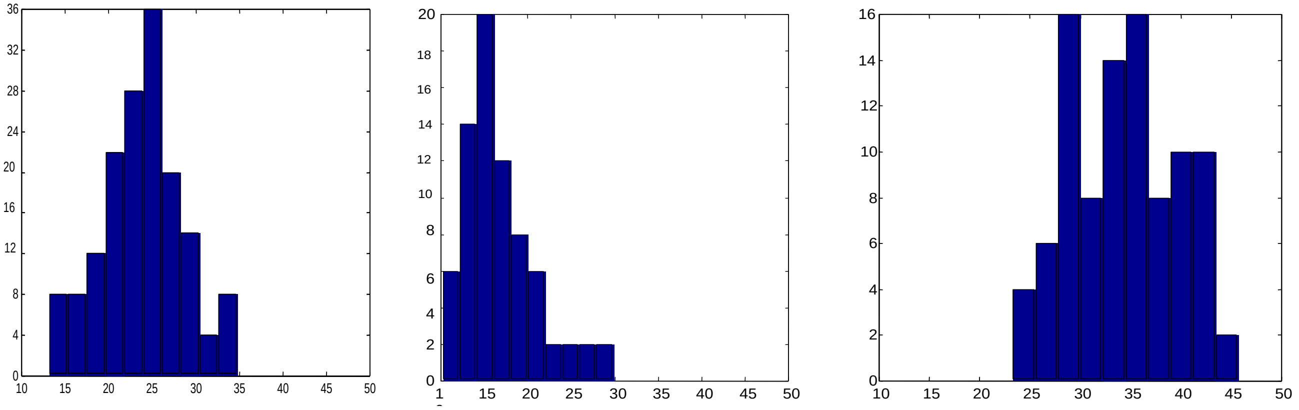
Classification of images drawn by the painters and those of the natural pictures have been achieved on the basis of the ratio of the area containing the dominant colors to that of the total area from their respective hue histograms. The percentage r_a of the area bounded by the five dominant colors to that of the total area has been calculated from the hue histograms for different images. The hue component values of five dominant colors found for different actors is given in Table 1. For natural pictures the dominant colors change within the range [0,360] depending on the color properties of the images.

TABLE I. HUE COMPONENT VALUE OF DIFFERENT ACTOR

Name of painters	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**
Sharaku	{61}	{91}	{121}	{181}	{27..39}
Hokusai	{27..41}	{241}	{91}	{181}	{61}
Hiroshige	{91}	{181}	{209..215}	{241}	{45..56}

The r_a versus the number of occurrences graph is shown in Fig. 5, which reveals that if the threshold value is taken up to 1.0σ, where σ is the variance, for the five dominant colors, pictures drawn by Sharaku will get the r_a value within the range [19.29,28.67], Hokusai within the range [12.45,20.77], and Hiroshige within [29.90,39.72], respectively. Relating Sharaku, there is an outstanding difference from other painters – almost all of his drawn pictures are of Ukiyoe actors.

The painters used only a few number of colors during drawing pictures. The number of colors used by different painters has been justified for different threshold values of r_a and is shown in Fig. 6, which reveals that the natural pictures have got innumerable number of colors whereas those of the painting pictures are limited. Finally, the Ukiyoe actors are distinguished from the normal pictures according to the RGB color distribution in the face color. Fig. 7 shows the skin color distribution of 80 Ukiyoe actors faces, where two distinct color zones are found. For the larger cluster, mean values for red, green and blue are m_R =103.65, m_G =104.13, m_B =104.14, and the variances are σ_R =39.07, σ_G =37.84, σ_B =40.15, respectively.



(a) Sharaku (μ=23.98, σ=4.69)

(b) Hokusai (μ=16.61, σ=4.16)

(c) Hiroshige (μ=34.31, σ=5.41)

Fig. 5. r_a versus number of occurrences

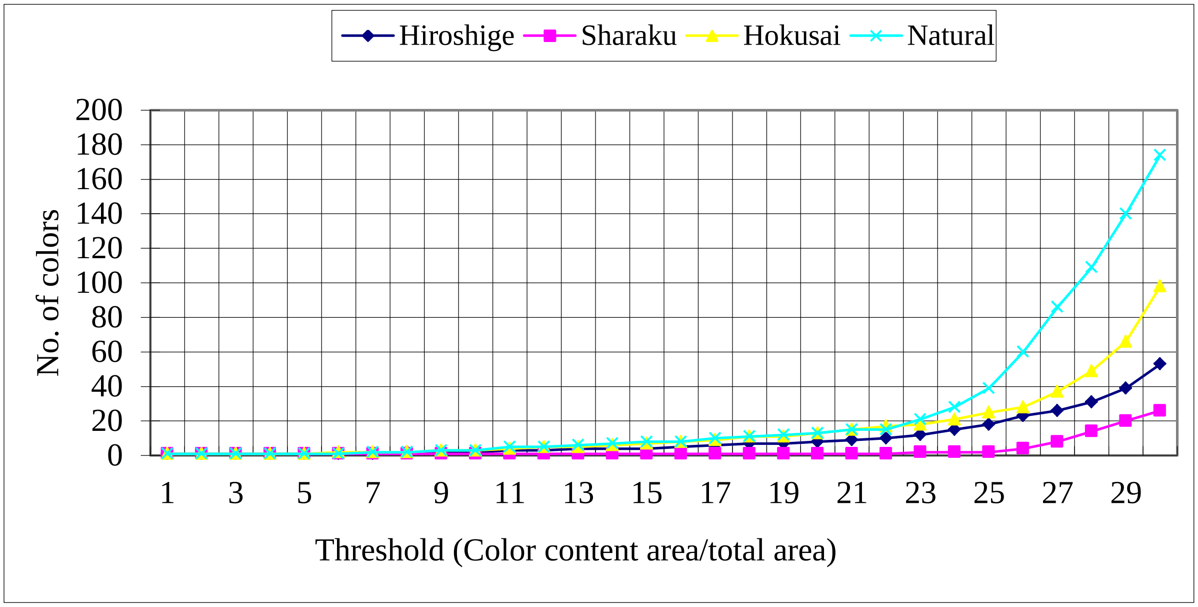
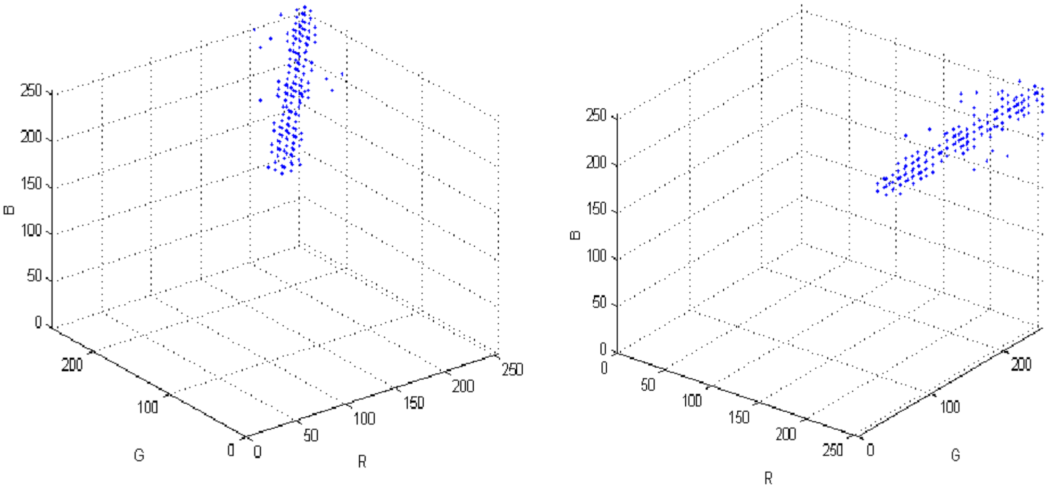


Fig. 6. r_a versus number of colors



(a) View (azimuth=37.5°, elevation=30°)

(b) View (azimuth=-37.5°, elevation=30°)

Fig. 7. Skin color cluster and distribution for 80 Ukiyoe faces

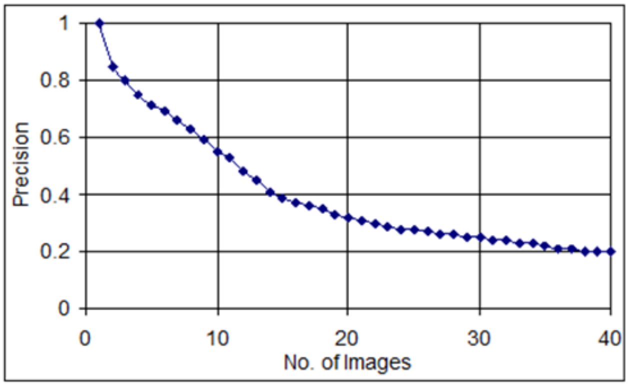


Fig. 8. Precision versus No. of Images

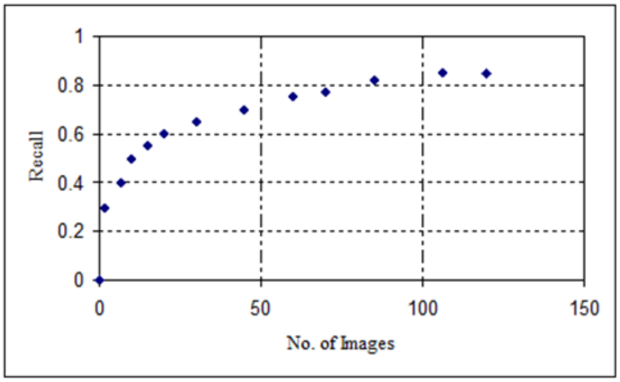


Fig. 9. Recall versus No. of Images

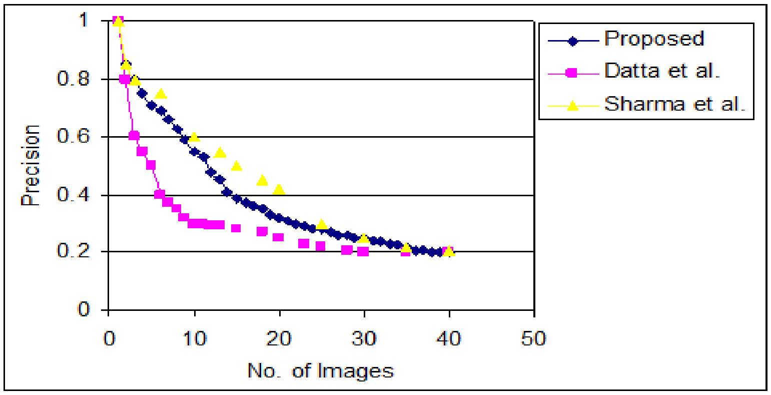


Fig. 10. Precision versus No. of Images for different methods

The image retrieval system has been accessed by two commonly used evaluation measures: (i) precision and (ii) recall, as defined by:

$$\text{Precision} = \frac{\text{No. of relevant images retrieved}}{\text{Total No. of images retrieved}} \qquad (11)$$

$$\text{Recall} = \frac{\text{No. of relevant images retrieved}}{\text{Total No. of images in the collection}} \qquad (12)$$

The precision versus number of images and recall versus number of images are shown graphically in Fig. 8 and Fig. 9, respectively, for different classes of images. Finally, the proposed method has been compared with the existing influential methods where similar measures have been established. The comparison has been made in terms of precision and is shown graphically in Fig. 10. The graph reveals that the proposed method performs better for less number of images and for higher number of images the performance is more or less the same like existing methods.

VII. Colonclusion

This paper describes the design and implementation of the content-based image retrieval system. So a quantitative analysis of color distribution has been presented for searching, indexing and retrieving color images. In the query process, the goal of the query is to retrieve images of interest. Prior to the trials, 450 images has been inspected, among them 320 are designated as painting pictures according to the ratio of the dominant colors to those of the total areas from the hue histograms. When the query is issued, the corresponding color index file is analyzed to select a set of candidate images containing regions with the similar colors of the query. The major limitation of the proposed method is that similarity measure for image retrieval has been established on the basis of Minkowski distance metric with L_2 norm. Other types of similarity measures like Mahalanobis metric or hausdorff distance can be considered in future for expressing similarity between colors. Our future plan is to develop a multimedia information system that will be able to perform the storage, browsing, indexing, and retrieval of multimedia data based on their text, sound and video contents.

Acknowledgment

The author would like to express his profound gratitude to Late Professor Susumi Matsudaira of Sonoda Womens' University, Japan for letting him use his enthusiastically collected Ukiyoe pictures. He is thankful to Professor Hiromitsu Hama of Osaka City University for his pragmatic suggestions regarding to this research work. The author is also grateful to all of the WWW users whose collected images enhanced his research work as a tool for experimental verification.

References

[1] Y. Liu, D. Zhang, G. Lu, and W. Ma, "A survey of content-based image retrieval with high-level semantics", Pattern Recognition, Vol. 40, No. 1, pp. 262-282, 2007.

[2] F. Long, H. Zhang, and D.D. Feng, "Fundamentals of Content based Image Retrieval", Multimedia Information Retrieval and Management, Springer, pp. 1-26, 2003.

[3] T. Deselaers, D. Keysers, and H. Ney, "Features for Image Retrieval: An Experimental Comparison", Information Retrieval, Vol. 11, No. 2, pp. 77-107, 2008.

[4] R. Datta, D. Joshi, J. Li, and J.Z. Wang, " Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, Vol. 40, No. 2, pp. 5:1-5:59, 2008.

[5] Md. Al-Amin Bhuiyan and Hiromitsu Hama, "Identification of Actors Drawn in Ukiyoe Pictures", Pattern Recognition, Vol. 35, No. 1, pp. 93-102, 2002.

[6] T. Gevers and A.W.M. Smeulders, "Content based image retrieval by viewpoint-invariant color indexing", Image and Vision computing, Vol. 17, No. 7, pp. 475-488, 1999.

[7] I.K. Park, I.D. Yun, and S.U. Lee, "Color image retrieval using hybrid graph representation", Image and Vision Computing, Vol. 17, No. 7, pp. 465-474, 1999.

[8] R.K. Srihari, "Automatic indexing and content-based retrieval of captioned images", IEEE computer, Vol. 28, No. 9, pp. 49-56, 1995.

[9] A.K. Jain and A. Vailaya, "Image retrieval using color and shape", Pattern Recognition, Vol. 29, No. 8, pp. 1233-1244, 1996.

[10] J. Martinez and S. Guillaume, "Color image retrieval fitted to classical querying", Proc. ICIAP II, pp. 14-21, 1997.

[11] A.D. Bimbo, M. Mugnaini, P. Pala, F. Turco, and L. verzucoli, "Image retrieval by color regions", Proc. ICIAP II, pp. 180-185, 1997.

[12] X. Wan and C.C. Jay kuo, Color distribution analysis and quantization for image retrieval", Storage and Retrieval for Image and Video Databases IV, SPIE 2670, pp. 8-16, 1995.

[13] T.F. Sayeda-Mahmood, "Data and model driven selection using color regions", International Journal of Computer Vision, Vol. 21, No. 1, pp. 9-36, 1997.

[14] Pentland, R.W. Picard, and S. Sclaroff, Photobook: Tools for content-based manipulation of image databases", Storage and Retrieval for Image and Video Databases II, SPIE 2185, 1994.

[15] C. Colombo, A. Rivi, and I. Genovesi, "Histogram families for color-based retrieval in image databases", Proceedings of ICIAP II, pp. 204-211, 1997.

[16] T. Caelli, D. Reye, "On the classification of image regions by color, texture and shape", Pattern Recognition, Vol. 26, No. 4, pp. 461-470, 1993.

[17] M.J. Swain and D.H. Ballard, "Color indexing", Int. J. Computer Vision, Vol. 7, No. 1, pp. 11-32, 1991.

[18] R. Kieldsen and J. Kender, "Finding skin in color images", 2nd Int. Conf. Automatic Face Gesture Recognition, 1996.

[19] B.V. Funt and G.D. Finlayson, "Color constant color indexing", IEEE Trans. Pattern Anal. Machine Intel, Vol. 17, No. 5, pp. 522-529, 1995.

[20] F. Ennesser, G. Medioni," Finding Waldo, or Focus of Attention Using Local Color Information", IEEE Trans. Pattern Anal. Machine Intel, Vol. 17, No. 8, pp. 805-809, 1995.

[21] C.C. Chang and L.L. Wang, "Color texture segmentation for clothing in a computer-aided fashion design system", Image Vision and Computing, Vol. 14, No. 9, pp. 685-702, 1996.

[22] S.J. McKenna, Y. Raja and S. Gong, "Tracking color objects using adaptive mixture models", Image and Vision Computing, Vol. 17, No. 3, pp. 225-231, 1999.

[23] D. Androutsos, K.N. Plataniotis, and A.N. Venetsanopoulos, "A novel vector-based approach to color image retrieval using a vector angular-based distance measure", Computer Vision and Image Understanding, Vol. 75, No. 1, pp. 46-57, 1999.

[24] Y. Liu and S. Ozawa, "A new representation and detection of multi-colored object based on color contents", IEICE Trans. Info. System, Vol. E83-D, No. 5, pp. 1170-1176, 2000.

[25] N. Sharma, P. Rawat and J. Singh, "Efficient CBIR using color histogram processing" Signal & Image Processing: An International Journal (SIPIJ), Vol.2, No.1, pp. 94-112, 2011.

[26] Md. Al-Amin Bhuiyan, Vuthichai Ampornaramveth, Shin-yo Muto, and Haruki Ueno, "On Tracking of Eye for Human-robot Interface", International Journal of Robotics and Automation, Vol. 19, No. 1, pp. 42-54, 2004.

[27] Md. Al-Amin Bhuiyan, Chang Hong Liu and Haruki Ueno, "On Pose Estimation for Human-Robot Symbiosis", International Journal of Advanced Robotic Systems, Vol. 5, No. 1, pp. 19-30, 2008.

[28] Md. Al-Amin Bhuiyan, Vuthichai Ampornaramveth, Shin-yo Muto, and Haruki Ueno, "Eye Tracking and Gaze Direction for Human-robot Interaction", 7th Rototics Symposia, Japan Robotics Society, pp. 209-214, 2002.

[29] A.R. Weeks, L.J. Sartor, and H.R. Myler, "Histogram specification of 24-bit color images in the color difference (C-Y) color space", J. Electronic Imaging, Vol. 8, No. 3, pp. 290-300, 1999.

[30] V. Buzuloiu, M. Ciuc, R.M. Rangayyan, and C. Vertan, "Adaptive-neighborhood histogram equalization of color images", 10(2), 445-459 (2001).

[31] C. Colombo and A.D. Bimbo, "Color-induced image representation and retrieval", Pattern Recognition, Vol. 32, No. 10, pp. 1685-1695, 1999.

[32] J. Yang, W. Lu, A. Waibel, "Skin-color modeling and adaptation", Proc. ACCV, pp. 687-694, 1998.

[33] T. Yoo and I. Oh, "A fast algorithm for tracking human faces based on chromatic histograms", Pattern Recognition Letters, Vol. 20, No. 10, pp. 967-978, 1999.

[34] Y.M. Lim and S.U. Lee, "On the color image segmentation algorithm based on thresholding and fuzzy c-means techniques, Pattern Recognition, Vol. 23, No. 9, pp. 935-952, 1990.

Multiple-Published Tables Privacy-Preserving Data Mining: A Survey for Multiple-Published Tables Techniques

Abou_el_ela Abdo Hussein
Department of Computer Science
Faculty of Science and Arts,
Shaqra University
Shaqra, KSA

Nagy Ramadan Darwish
Sciences, Institute of Statistical
Studies and Research, Cairo
University,
Cairo, Egypt

Hesham A. Hefny
Department of Computer and
Information Sciences, Institute of
Statistical Studies and Research,
Cairo University,
Cairo, Egypt

***Abstract*—With large growth in technology, reduced cost of storage media and networking enabled the organizations to collect very large volume of information from huge sources. Different data mining techniques are applied on such huge data to extract useful and relevant knowledge. The disclosure of sensitive data to unauthorized parties is a critical issue for organizations which could be most critical problem of data mining. So Privacy preserving data mining (PPDM) has become increasingly popular because it solves this problem and allows sharing of privacy sensitive data for analytical purposes. A lot of privacy techniques were developed based on the k-anonymity property. Because of a lot of shortcomings of the k-anonymity model, other privacy models were introduced. Most of these techniques release one table for research public after they applied on original tables. In this paper the researchers introduce techniques which publish more than one table for organizations preserving individual's privacy. One of this is (α, k) – anonymity using lossy-Join which releases two tables for publishing in such a way that the privacy protection for (α, k)-anonymity can be achieved with less distortion, and the other one is Anatomy technique which releases all the quasi-identifier and sensitive values directly in two separate tables, met l-diversity privacy requirements, without any modification in the original table.**

Keywords—Data mining; privacy; sensitive attribute; quasi-identifier; Anatomy

I. Introduction

One of the most important aspects of data applications is data mining. Data mining technique intelligently and automatically extracts information or knowledge from a very large volume of data.

One of the disadvantages of data mining is the disclosure of sensitive individual data to unauthorized parties which are a critical issue for organizations. So Privacy Preserving Data Mining (PPDM) is playing very important role in both applications and research; it publishes much more accurate data while maintaining privacy information. Each record in released data corresponding to one individual and has a number of attributes, which can be divided into three categories:

*1) **Identity attributes** (e.g., SSN and Name) whose values can uniquely identify an individual;*

*2) **Quasi-identifier** (QI-group) attributes (e.g., age, Zip code and gender) whose values can potentially identify an individual;*

*3) **Sensitive attributes** (e.g., income and disease) which indicate confidential and sensitive information of individuals.*

Several Privacy-Preserving data mining techniques have been published most of them depending on k-anonymity. *The anonymization techniques* (e.g. *k-anonymity)* aim at using techniques of generalization and suppression to make the individual record indistinguishable from a group of records. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers that are used with public records to uniquely identify the records. Because of k-anonymity has some shortcomings, many advanced methods have been proposed, such as *p*-sensitive k-anonymity, (*α, k*)-anonymity, *l*-diversity, *t*-closeness, *M*-invariance, Personalized anonymity, and so on. Although the anonymization method can ensure that the transformed data is true, it also results in information loss to some extent [1]. Also, there is a technique called Anatomy technique that releases all the quasi-identifier and sensitive values directly in two separate tables. In this paper, researchers focus only on those techniques that publish more than one table for the purposes of data mining. In next section, researchers introduce k-anonymity technique and both generalization and suppression concepts. In section three both multiple-published tables techniques, Anatomy and (α, k) – anonymity using lossy-Join ending with a comparison between them are introduced, and last section introduces paper conclusion.

II. Related Research Areas

Numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression. First the researchers introduce K-anonymity Technique proposed by L. Sweeney in next sub-section, then generalization and suppression concepts are introduced in last sub-section.

A. K-anonymity Technique

K-anonymity classified the attributes into three classes as mentioned before [2]. Table I. introduces the three classes of attributes where, ***Identity attributes*** (e.g., Name), ***Quasi-***

identifier (*QI-group*) attributes (e.g., gender, age and Zip code), ***Sensitive attributes*** (e.g., Diagnosis). K-anonymity technique anonymizes *QI-group* to prevent the attacker using link attack to infer the privacy of individuals. *Quazi-identifiers* can be used to re-identify individual using linking attack as given in below example.

TABLE I. CLASSIFICATION OF ATTRIBUTES FOR K-ANONYMITY

Identifier attribute	Quasi-identifier			Sensitive attributes
Name	Gender	Age	Zip code	Diagnosis
Ali	Male	25	423101	Depression
Mohsen	Male	27	423508	HIV

The two tables, Table II. contains Medical data set and Table III. Contains voter list which are available publically. To avoid the identification of records in microdata, the traditional approach is to de-identify records by removing the ***identity attribute*** (e.g., Name). But removing the ***identity attribute*** does not solve the problem because by linking Zipcode, Age and Sex of medical table (Table II.) with voter list table (Table III.) intruder can disclose that Omar is sick with cancer and in this way the privacy of individual is disclosed. This is happened because the combination of quazi-identifiers value is unique in medical data set, if published data in such a way that there is no unique combination for quazi-identifiers then this type of re-identification cannot occurs. This can be done using anonymizing tables.

TABLE II. MEDICAL DATA SET

ID	Zip code	AGE	SEX	DIAGNOSIS
1	423065	29	M	Heart Disease
2	422036	32	F	Flu
3	423245	38	M	Cancer
4	422035	37	F	HIV
5	423012	47	M	Headache
6	423432	53	F	Viral

Sweeney [1] proposed the k-anonymity model in order to prevent linking attacks using quasi-identifiers, where some of the *QI* fields are generalized or suppressed. A table is said to satisfy k-anonymity if every record in the table is indistinguishable from at least k-1 other records to every set of quasi-identifier attributes. The table is called a k-anonymous table if, for every combination of attributes of the *QI*s, there are at least k records that share those values. This ensures that individuals cannot be uniquely identified using linking attacks. Table IV. shows a 2-anonymous view corresponding to Table II. The sensitive attributes (Diagnosis Result) is stayed without change in this example.

TABLE III. VOTER LIST

NAME	Zipcode	AGE	SEX
Mohamed	423234	49	M
Ahmed	466987	35	M
Ali	423223	28	M
Rawan	424435	41	F
Omar	423245	38	M
Iman	423446	33	F

TABLE IV. 2-ANONYMOUS VIEW OF TABLE II.

ID	Zipcode	AGE	SEX	DIAGNOSIS
1	423***	>25	M	Heart Disease
2	423***	>25	M	Cancer
3	422***	3*	F	Flu
4	422***	3*	F	HIV
5	423***	>40	*	Headache
6	423***	>40	*	Viral

Numerous techniques implementing k-anonymity have been proposed using generalization and suppression [3]. Generalization involves modifying (or recoding) a value with a less specific but semantically consistent value. Suppression involves not publishing a value at all. An algorithm that exploits a binary search on the domain generalization hierarchy to find minimal k-anonymous table have been proposed by Samarati [4]. A. Machanavajjhala [5] proposed l-diversity technique in 2006 to solve k-anonymity problem. It tries to put constraints on minimum number of distinct sensitive values seen within an equivalence class , T-closeness technique present by S. Venkatasubramanian in 2007 [6] to overcome attacks possible on l-diversity like similarity attack[7], Bayardo and Agrawal [8] presented technique that starts from a fully generalized table and specializes the dataset in a minimal k- anonymous table. R. Wong, J. Li, A. Fu, K. Wang [9] proposed an (α, k)-anonymity technique to protect both identifications and relationships to sensitive information in data in the literature in order to deal with the problem of k-anonymity. Fung et al. [10] presented a top-down approach to make a table satisfied k-anonymous. LeFevre et al [11] introduces technique that uses a bottom-up technique. Pei [12] discusses the approaches for multiple constraints and incremental updates in k-anonymity. However the traditional k-anonymity techniques take consider that the all values of the sensitive attributes are sensitive and need to be protected. The previous models lead to excessively generalize and more information loss in publishing data.

B. Generalization and Suppression

Generalizing an attribute is a simple concept idea. A value is replaced by a less specific, more general value that is faithful to the original [1, 13, 14, and 15]. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. Generalization could be achieved through global recoding or local recoding. In global recoding, the domain of the quasi identifier values are mapped to generalized values for achieving k-anonymity, which means that all k-tuples have the same generalized attribute value.

In local recoding generalization scheme, any two or more regions can be merged as long as the aggregated attribute value such as satisfies the anonymity requirement, which means that each k-tuple could have its own generalization attribute value. The limitation of the global recoding is; the domain values are over generalized resulting in utility loss where as in local recoding, the individual tuple is mapped to a generalized tuple.

The information loss of the global recoding is more than the local recoding approach. Comparison between global and local recoding is in table V.

TABLE V. COMPARISON BETWEEN GLOBAL AND LOCAL RECODING

Generalization method	**Global Recoding**	**Local Recoding**
Generalized value	The same value	Different values
Information loss	More information loss	Less information loss
Utility level	More utility loss	Less utility loss
Domain values	Over generalized	Suitable generalization
Generalization kind	Global generalized	Local generalized
Privacy level	High level of privacy	Lower level of privacy
Generalization level	Higher generalization level	Minimum generalization level

While Generalization replaces the actual *QI* values with more general ones (e.g., replaces the city name with the state name); Suppression involves not releasing a value at all. Suppression is the most common practice in related works on such data. There is a generalization hierarchy (e.g., city name→ state name → country name). On the other hand Suppression excludes some *QI* attributes or entire records (known as outliers) from the microdata. Comparison between generalization and suppression in table VI.

TABLE VI. COMPARISON BETWEEN GENERALIZATION AND SUPPRESSION

Method	Generalization	Suppression
Generalized value	Releases general value	Not releasing value at all
Information loss	Less information loss	More information loss
Utility level	Less utility loss	More utility loss
Privacy level	Lower privacy level	High privacy level
Common method	Less common	More common practice

Both Generalization and Suppression Architecture could be explained by figure 1.

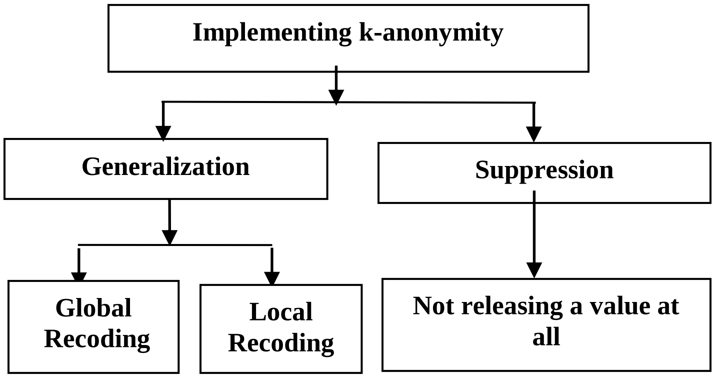


Fig. 1. generalization and suppression

III. MULTIPLE-PUBLISHED TABLES

In this section the researchers introduce techniques which publish more than one table for organizations preserving individual's privacy. One of this is (α, k) – anonymity using lossy-Join which releases two tables for publishing and the other one is Anatomy technique which releases all the quasi-identifier and sensitive values directly in two separate tables. Next subsections introduce these two techniques in details.

A. (α, k) – anonymity using lossy-Join

The Lossy Join Approach

Lossy join of multiple tables is useful in privacy-preserving data publishing [9]. The mean idea is that if two tables with a join attribute are released, the join of the two tables can be lossy and this lossy join helps to maintain the private information. In this paper, authors use the idea of lossy join to derive a new technique for achieving privacy preservation purpose. Let us see Table VII. an (0.5, 2) -anonymization. From this table, temp table could be generated as shown in Table VIII.

For each equivalence class *E* in the anonymized table, there is a unique identifier (*ID*) to *E* and also to all tuples in *E*. Then, the correspondence *ID* to each record in the original raw table could be attached and form a new table called Temp. From the Temp table, two separate tables could be generated, Tables IX.(a) and IX.(b). The two tables share the attribute of *ClassID*. If these two tables are joined using the *ClassID*, the join is lossy and it is not possible to obtain the table Temp after the join. The resulted table is given in Table X.

From the lossy join, each individual is linked to at least 2 values in the sensitive attribute. Therefore, the required privacy of individual can be maintained.

Also, in the joined table, for each individual, there are at least 2 persons that are linked to the same bag *B* of sensitive values, so they are not distinguishable.

For example, the first record in the raw table (*QID* = (clerk, 1975, 4350)) is linked to bag {*HIV*, *flu*}. The second record (*QID* = (manager, 1955, 4350)) is also linked to the same bag *B* of sensitive values. This is the goal of k-anonymity for the protection of sensitive values.

TABLE VII. AN (0.5, 2)-ANONYMOUS DATA SET

Job	**Birth**	**Post Code**	**Illness**
Clerk	1975	4350	HIV
manager	1955	4350	flu
clerk	1955	5432	flu
factory worker	1955	5432	fever
factory worker	1975	4350	flu
technical supporter	1940	4350	fever

TABLE VIII. TEMP TABLE

Job	**Birth**	**Post Code**	**Illness**	**ClassID**
Clerk	1975	4350	HIV	1
manager	1955	4350	flu	1
clerk	1955	5432	flu	2
factory worker	1955	5432	fever	2
factory worker	1975	4350	flu	3
technical supporter	1940	4350	fever	3

B. *Anatomy*

Anatomy releases two different tables *QI* (*Quisi-identifier)* attributes table and *SI* (*Sensitive)* attributes table instead of publishing one single table with the generalized values. Anatomy [16] releases all *QIs* and *SI* directly in two separate tables, which met *L*-diversity privacy requirement, so there is no need to modify the original table. Anatomy avoids the drawbacks of generalization as in next example. Assume that hospital intents to publish patients' medical records as in Table XI., referred to as the microdata.

The sensitive Attribute is Disease, so the hospital must ensure that no intruder can correctly infer any patient disease with confidence. Age, Sex, and Zipcode are the quasi-identifier (*QI*) attributes, which could be utilized in combination to infer the identity of an individual, which disclose privacy.

TABLE IX. (A): NSS TABLE

Job	**Birth**	**Post Code**	**ClassID**
Clerk	1975	4350	1
manager	1955	4350	1
clerk	1955	5432	2
factory worker	1955	5432	2
factory worker	1975	4350	3
technical supporter	1940	4350	3

(B): SS TABLE

ClassID	**Illness**
1	HIV
1	flu
2	flu
2	fever
3	flu
3	fever

Consider an intruder who has the personal details (i.e., age 25 and Zipcode 11500) of Ali, and knows that Ali has been hospitalized before. In Table XI., since only record 1 matches Ali's QI-values, the adversary knows that Ali has pneumonia. To avoid this problem, generalization [4, 17, 18, and 5] divides records into *QI-groups*, and transforms their *QI-values* into less specific forms, so that records in the same *QI-group* cannot be distinguished by their *QI-values*. Table XII. is a generalized version of Table XI. (e.g., the age 25 and Zipcode 11500 of record 1 have been replaced with intervals [19, 20] and [10001, 60000], respectively). Here, generalization produces two QI-groups, including records 1-4 and 5-8, respectively. As a result, even if an intruder has the exact QI values of Ali, s/he still does not know which record in the first QI-group belongs to Ali.

Two notions, k-anonymity and l-diversity, have been proposed to measure the degree of privacy preservation. A (generalized) table is k-anonymous [4, 17, 18] if each QI-group involves at least k records (e.g., Table XII. is 4-anonymous). However, even with a large k as shown in *l-diversity* [5], k-anonymity may still allow an intruder to infer the sensitive value of an individual with high confidence. So, *l-diversity in* [5] provides stronger privacy preservation.

TABLE X. SS TABLE

Job	Birth	Post Code	Illness	ClassID
Clerk	1975	4350	HIV	1
manager	1955	4350	HIV	1
Clerk	1975	4350	flu	1
manager	1955	4350	flu	1
clerk	1955	5432	flu	2
factory worker	1955	5432	flu	2
clerk	1955	5432	fever	2
factory worker	1955	5432	fever	2
factory worker	1975	4350	flu	3
technical supporter	1940	4350	flu	3
factory worker	1975	4350	fever	3
technical supporter	1940	4350	fever	3

TABLE XI. THE MICRODATA

Tuple ID	Age	Sex	Zipcode	Disease
1(Ali)	25	M	11500	pneumonia
2	29	M	13200	dyspepsia
3	33	M	59300	dyspepsia
4	55	M	12700	pneumonia
5	60	F	54600	flu
6	59	F	25200	gastritis
7(Hoda)	60	F	25100	flu
8	58	F	31000	bronchitis

Specifically, a table is ***l***-diverse if, in each *QI-group*, at most 1/***l*** of the records possesses the most frequent sensitive value1. For instance, Table XII. is 2-diverse because, in each *QI-group*, at most 50% of the records have the same value of Disease. As mentioned earlier, the intruder (targeting Ali's medical record) knows that Ali's record must be in the first

QI-group, where two records are associated with pneumonia, and two with dyspepsia. Hence, the adversary can only make a probabilistic conjecture: Ali could have either disease with the same probability.

TABLE XII. A 2-DIVERSE TABLE

Tuple ID	Age	Sex	Zipcode	Disease
1	[21, 60]	M	[10001, 60000]	pneumonia
2	[21, 60]	M	[10001, 60000]	dyspepsia
3	[21, 60]	M	[10001, 60000]	dyspepsia
4	[21, 60]	M	[10001, 60000]	pneumonia
5	[21, 60]	F	[10001, 60000]	flu
6	[21, 60]	F	[10001, 60000]	gastritis
7	[21, 60]	F	[10001, 60000]	flu
8	[21, 60]	F	[10001, 60000]	bronchitis

Anatomy technique has been proposed to overcome the disadvantages of generalization which often losses considerable information in the microdata. Anatomy captures the exact *QI*-distribution and releases two tables, a quasi-identifier table (*QIT*) and a sensitive table (*ST*), which separate *QI*-values from sensitive values. For example, Tables XIII.(a) and XIII.(b) demonstrate the *QIT* and *ST* obtained from the microdata Table XI., respectively [16].

First, the microdata partitioned the records into different QI-groups, based on a certain strategy. Here, following the grouping in Table XII., records 1-4 into *QI*-group number 1and records 5-8 into QI-group number 2 of Table XI.

Second, the quasi-identifier table (*QIT*) has been created. Specifically, for each record in Table XI., the *QIT* (Table XIII.(a) includes all its exact *QI-values*, together with its group membership in a new column Group-ID. However, *QIT* does not have any Disease value.

Finally, it is possible saying that ST (Table XIII.(b) maintains the Disease statistics of each *QI*-group.

Anatomy preserves privacy because the *QIT* does not indicate the sensitive value of any record, which must be randomly guessed from the ST. To explain this, consider again the adversary who has the age 25 and Zip code 11500 of Ali. Hence, from the *QIT* (Table XIII.(a), the adversary knows that record 1 belongs to Ali, but does not obtain any information about his disease so far. Instead, s/he gets the id 1 of the QI-group containing record 1. Judging from the ST (Table XIII.(b), the adversary realizes that, among the 4 records in QI-group 1, 50% of them are associated with pneumonia (or dyspepsia) in the micro data. Note that s/he does not gain any additional information, regarding the exact diseases carried by these records. Hence, s/he could only expect that Ali could have contracted pneumonia (or dyspepsia) with 50% probability.

TABLE XIII. THE ANATOMIZED TABLES

(a) The quasi-identifier table (QIT)

row #	Age	Sex	Zipcode	Group-ID
1(Ali)	25	M	11500	1
2	29	M	13200	1
3	33	M	59300	1
4	55	M	12700	1
5	60	F	54600	2
6	59	F	25200	2
7(Hoda)	60	F	25100	2
8	58	F	31000	2

(b) The sensitive table (ST)

Group-ID	Disease	Count
1	Dyspepsia	2
1	Pneumonia	2
2	Bronchitis	1
2	Flu	2
2	Gastritis	1

Researchers introduce Comparison between Anatomy [16] and (α, k) – anonymity using lossy-Join [9] in table XIV.

TABLE XIV. COMPARISON BETWEEN ANATOMY AND (A, K) – ANONYMITY USING LOSSY-JOIN

Technique	**Anatomy**	**(α, k) – anonymity using lossy-Join**
No. of tables	Two tables	Two tables
l diverse	Achieve *l* diversity	Achieve *l* diversity
Information Loss	No Information Loss	There is Information Loss
Data Utility	More Data Utility	Less Data Utility

The Architecture for both Multi-published tables' techniques is represented in figure 2.

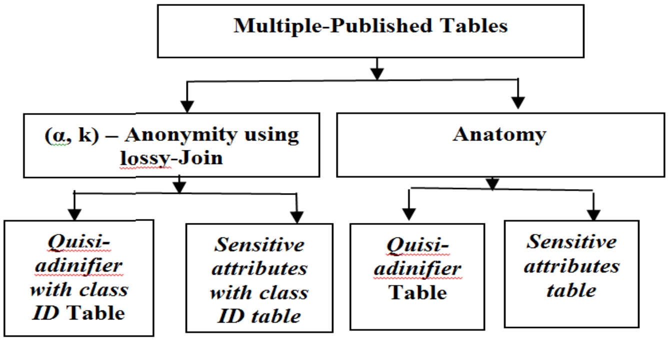

Multiple-Published Tables
(α, k) – Anonymity using lossy-Join
Anatomy
Quisi-adinifier with class ID Table
Sensitive attributes with class ID table
Quisi-adinifier Table
Sensitive attributes table

Fig. 2. multi-published tables archetictcher

IV. Conclusion and Future Work

This paper introduces a survey for most common privacy preserving data mining techniques that using Multiple-Published Tables PPDM & PPDP and explains their effects on Data Privacy. Both Anatomy and (α, k) – anonymity using lossy-Join [9] are used for security of respondents identity and decreases linking attack. It is observed that using generalization and suppression in (α, k) – anonymity using lossy-Join technique on those attributes lead to reduce the precision of publishing table. (α, k) – anonymity using lossy-Join also causes data lose because suppression emphasize on not releasing values which are not suited for k factor although it maintaining privacy. The idea of (α, k) – anonymity using lossy-Join is that if two tables with a join attribute are published, the join of the two tables can be lossy that helps to maintain the private information. On the other hand anatomy technique applied on sensitive tables reduces information loss, because it releases all the quasi-identifier and sensitive values directly in two separate tables without applying any suppression or even any generalization leads to data utility maintaining. The idea of Anatomy preserving privacy is that *QIT* does not indicate the sensitive value of any record, which is randomly guessed. Future work can include defining a new privacy technique for multiple sensitive attributes and researchers will focus to publish attributes without suppression using generalization boundaries technique that used to achieve k-anonymity maintaining individual privacy without influence data utility.

References

[1] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal on Uncertainty, Vol. 10, No. 5, pp. 571–588, 2002.

[2] Abou el ela A. Hussien, Nermin Hamza, Ashraf A. Shahen and Hesham A. Hefny, "A survey of privacy preserving data mining algorithms", Yanbu Journal of Engineering and Science, Vol. 5, October, 2012.

[3] Bhavana Abad (Khivsara), Kinariwala S.A., " A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity", International Journal of Computer Applications , Vol. 42, No.4, March, 2012.

[4] P. Samarati. "Protecting respondents identities in microdata release", IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No 6, 2001.

[5] Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-Diversity: Privacy beyond k-anonymity", In Proceedings of the IEEE ICDE, 2006.

[6] N. Li, T. Li, S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", ICDE, pp. 106-115, 2007.

[7] Abou el ela A. Hussien, Nermin Hamza, Hesham A. Hefny, "Attacks on Anony-mization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing", Journal of Information Security jis, vol 4, pp.101-112, April, 2013.

[8] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymity", In Proceedings of the 21st International conference on Data Engineering (ICDE), pp. 217-228, Tokyo, Japan, 2005.

[9] Raymond Chi-Wing Wong, Yubao Liu, Jian Yin, Zhilan Huang, AdaWai-Chee Fu1, and Jian Pei, " (α, k)-anonymity Based Privacy Preservation by Lossy join", Lecture Notes in Computer Science,Vol. 4, pp. 733-744, 2007.

[10] B. Fung, K. Wang, P. Yu, "Top-down Specialization for Information", Conference on Data Engineering (ICDE), pp. 205-216, 2005.

[11] K LeFevre, D DeWitt, R Ramakrishnan. Incognito," Efficient full domain k-anonymity", Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 49-60, Baltimore, Maryland, 2005.

[12] J Pei, J Xu, Z. B Wang, W Wang, K Wang, " Maintaining k- anonymity against incremental updates", Proceedings of the 19th International Conference on Scientific and Statistical Database, 2007.

[13] K. Venkata Ramana and V.Valli Kumari, "Graph Based Local Recoding For Data Anonymization", International Journal of Database Management Systems (IJDMS), Vol.5, No.4, August, 2013.

[14] Jian Xu Wei Wang Jian Pei Xiaoyuan Wang Baile Shi Ada Wai-Chee Fu3 "Utility-Based Anonymization Using Local Recoding", KDD'06, Philadelphia, Pennsylvania, USA. pp.20-23, August, 2006.

[15] Manolis Terrovitis _ Nikos Mamoulis _ Panos Kalnis," Local and Global Recoding Methods for Anonymizing Set-valued Data", Research Center Mathematical and Computer Sciences and Engineering,King Abdullah University of Science and Technology, 2011.

[16] Xiaokui, Xiao Yufei Tao, "Anatomy: Simple and Effective Privacy Preservation", VLDB, September, Seoul, Korea, 2006.

[17] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information". In PODS, page. 188, 1998.

[18] L. Sweeney, k-Anonymity: "A Model for Protecting Privacy", International Journal on Uncertainty, Vol. 10, No. 5, pp. 557–570, 2002.

[19] X. Xiao and Y. Tao. "Personalized privacy preservation", SIGMOD, 2006.

[20] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", In Proceedings of the 3rd International Conference on Data Mining, pp.99-106, 2003.

An Automated Graphical User Interface based System for the Extraction of Retinal Blood Vessels using Kirsch’s Template

Joshita Majumdar
Dept. of Electronics & Communication
IEM, WBUT
Kolkata, India

Debasish Kundu
Principle (ECE)
Gobindapur Sapahali Memorial Polytechnic
Burdwan, India

Souvik Tewary
Dept. of Bio Medical Engineering
NSEC, WBUT
Kolkata, India

Sudipta Ghosh
Dept. of Electronics & Communication
Bengal Institute of Technology & Management
Shantiniketan, India

Shreyosi Chakraborty
Dept. of Bio Medical Engineering
NSEC, WBUT
Kolkata, India

Sauvik Das Gupta
School of Electrical & Computer Engineering
Oklahoma State University
OK, USA

***Abstract*—The assessment of Blood Vessel networks plays an important role in a variety of medical disorders. The diagnosis of Diabetic Retinopathy (DR) and its repercussions including micro aneurysms, haemorrhages, hard exudates and cotton wool spots is one such field. This study aims to develop an automated system for the extraction of blood vessels from retinal images by employing Kirsch’s Templates in a MATLAB based Graphical User Interface (GUI). Here, a RGB or Grey image of the retina (Fundus Photography) is used to obtain the traces of blood vessels. We have incorporated a range of Threshold values for the blood vessel extraction which would provide the user with greater flexibility and ease. This paper also deals with the more generalized implementation of various MATLAB functions present in the image processing toolbox of MATLAB to create a basic image processing editor with different features like noise addition and removal, image cropping, resizing & rotation, histogram adjust, separately viewing the red, green and blue components of a colour image along with brightness control, that are used in a basic image editor. We have combined both Kirsch’s Template and various MATLAB Algorithms to obtain enhanced images which would allow the ophthalmologist to edit and intensify the images as per his/her requirement for diagnosis. Even a non technical person can manage to identify severe discrepancies because of its user friendly appearance. The GUI contains very commonly used English Language viz. Load, Colour Contrast Panel, Image Clarity etc that can be very easily understood. It is an attempt to incorporate maximum number of image processing techniques under one GUI to obtain higher performance. Also it would provide a cost effective solution towards obtaining high definition and resolution images of blood vessel extracted Retina in economically backward regions where costly machine like OCT (Optical Coherence Tomography), MRI (Magnetic Resonance Imaging) are not available. Hence an early detection of irregularity will be possible especially in rural areas.**

Keywords*—Blood Vessel Extraction; Retinal Images; Kirsch’s Templates; Image Processing; Graphical User Interface

I. Introduction

The vascular network is an essential anatomical structure in the human retina. An analysis of the retinal vasculature may lead to the diagnosis of various abnormalities such as haemorrhages, micro aneurysms etc. thus, an automated system for retinal blood vessel extraction is a preliminary step in the development of a computer-assisted diagnostic system for ophthalmic anomalies.

The extraction of the human eye vasculature from retinal images involves the essential tool of edge detection. According to Muthukrishnan & Radha [1], the classical methods of edge detection operates on convolving the image through an operator. Yin et al. [2] proposed a probabilistic tracking method to detect blood vessels in retinal images. They categorized the task of vessel extraction into two main groups: Pixel-based methods and tracking methods.

According to Zhang et al. [3], matched filter is a simple yet effective method for vessel extraction. They also proposed an extension to the matched filter approach to detect retinal blood vessels that significantly reduce the false detections produced by the original matched filter. Esmaeili et al. [4], presented an efficient algorithm for automatic extraction of blood vessels that comprises the following four steps: (i) Curvelet-based contrast enhancement, (ii) Matched filtering, (iii) Curvelet based edge-extraction and (iv) Length filtering. In their study, the enhanced image is first reconstructed from the modified curvelet co-efficient followed by match filtering to intensify the blood vessels along with the implementations of curvelet transform to segment vessels from its background. Finally, they have used length filtering to remove the misclassified pixels. Their experimental results have been evaluated on DRIVE database (Niemaiger & van Ginneken, [5]). The images on which we worked on have been collected from a publicly available DRIVE database and from the Regional Institute of Ophthalmology, Medical College, Kolkata [6].

The next section describes the hardware & software platforms of our system. Section III presents the methodology we proposed for blood vessel extraction. Sections IV & V deal with the relevant discussions and Future Prospects respectively. The conclusion of the paper is presented in Section VI.

II. Hardware and Software Platforms

Minimum Hardware requirements:

- Dual core processor
- RAM: 4 GB
- Windows XP/7/8
- 256 MB Graphics card(to enhance the performance of the GUI)

Our study has been conducted using:

- Laptop (SONY VAIO VPCEH16EN)
- Windows 7
- Intel® Core™ i3-2310M Processor 2.10 GHz
- RAM : 4 GB
- 256 MB nVIDIA GEFORCE® with CUDA.

The software used to develop the GUI is MATLAB which is a multi-paradigm numerical computing environment and 4th Gen. Programming language.

III. Methodology

The complete procedure of extracting the blood vessels from the coloured retinal image consist of three steps, beginning with fetching the input image from the system or the camera, then converting the RGB image into a gray scale image and finally using the kirsch's templates to detect the edges of the blood vessels. Fig. 1 illustrates the flow diagram of the proposed method for blood vessel extraction.

a) ***Fetching the image:*** *Two separate push buttons on the GUI facilitate the input of an image. One allows the user to browse an image from the system hard drive or any other external drives and the other allows the user to capture an image using any camera attached to the computer. The image that has been loaded is displayed on the first axis positioned on the top left corner of the GUI.*

b) ***Gray Conversion:*** *The panel 'Colour format' in the GUI contains a radio button 'GRAY Scale' that enables the user to convert the RGB retinal image to gray scale having pixel values ranging from 0 to 255 with just one mouse click.*

c) ***Extraction of Blood Vessels using Kirsch's Templates:*** *The Kirsch operator or Kirsch compass kernel is a non-linear edge detection that finds the maximum edge strength in a few pre determined directions. The 'VESSEL EXTRACTION' panel contains a popup menu with a number of threshold values (within a range of 4.6 to 9.0) for blood vessel extraction. The Kirsch operator can adjust the related threshold value automatically due to the image characteristics. The operator takes a single kernel mask and rotates it in 45° increments through all eight compass directions: North, North-West, West, South-West, South, South-East, East and North-East as shown below:*

$$M1=\begin{pmatrix}5&5&5\\-3&0&-3\\-3&-3&-3\end{pmatrix} M2=\begin{pmatrix}5&5&-3\\5&0&-3\\-3&-3&-3\end{pmatrix} M3=\begin{pmatrix}5&-3&-3\\5&0&-3\\5&-3&-3\end{pmatrix} M4=\begin{pmatrix}-3&-3&-3\\5&0&-3\\5&5&-3\end{pmatrix}$$

$$M5=\begin{pmatrix}-3&-3&-3\\-3&0&-3\\5&5&5\end{pmatrix} M6=\begin{pmatrix}-3&-3&-3\\-3&0&5\\-3&5&5\end{pmatrix} M7=\begin{pmatrix}-3&-3&5\\-3&0&5\\-3&-3&5\end{pmatrix} M8=\begin{pmatrix}-3&5&5\\-3&0&5\\-3&-3&-3\end{pmatrix}$$

The edge magnitude of the Kirsch operator is calculated as the maximum magnitude across all directions. Except the outermost rows and columns, every pixel along with its eight neighbouring pixels in a given image is convolved with the eight aforementioned templates respectively [7], providing eight outputs for each pixel, the maximum of which is defined as the edge magnitude (Gao et al. [8]. A pixel's Gray value with its eight neighbours is as shown.

P_0	P_1	P_2
P_7	P (i,j)	P_3
P_6	P_5	P_4

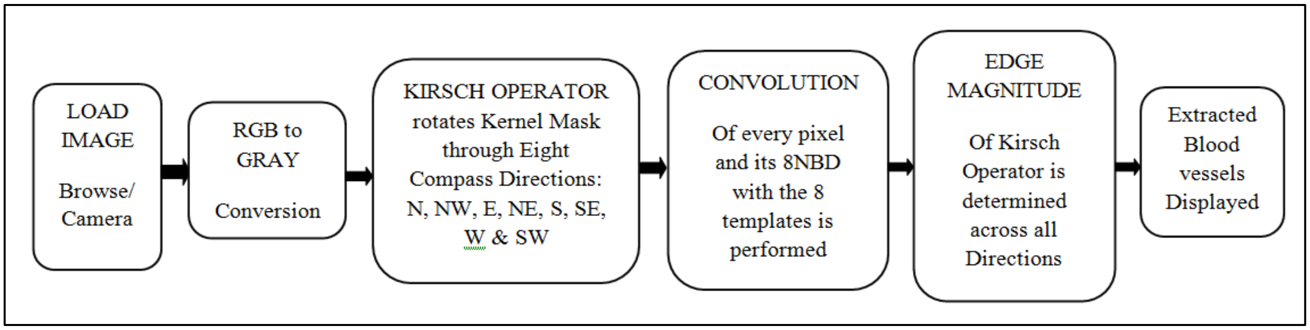


Fig. 1. Proposed Overall Method for Blood Vessel Extraction

The direction of edge is defined by the related mask that produces the maximum magnitude. The general output of edge detection through Kirsch's Templates is an image containing Gray level pixels of value 0 or 255. The value 0 indicates a black pixel and the value 255 indicates a white pixel. The edge information of the target pixel is checked by determining the brightness levels of the neighbouring pixels [9]. In case no major difference in the brightness level is found, the possibility of the pixel being a part of an edge is ruled out.

Kirsch's Template can set and reset the threshold values to obtain most suitable edge of images. It works well for images having a clear distinction between the foreground and background [10]. Since the retinal blood vessels can be considered as the required foreground information from the background fundus images, Kirsch's algorithm is effectively applicable. Fig. 2 illustrates an original fundus image with the extracted blood vessels.

d) ***Other Important Tools:***

A fact worth mentioning is that the GUI designed in order to carry out this study on the extraction of blood vessels may also be used to edit and enhance the retinal images for further analysis using the following options.

- Image Addition
- Image Adjust Panel
- Pixel Processing, Image Rotate, Image Complement, Colour Swap
- Image Resize, Image Crop
- Contrast Slider, R G B Sliders
- CMY Panel
- Colour Panel – Gray, B & W, Pseudo
- Noise Addition, Filter
- Edge Detection
- B & W Morphological Operations
- Colour Plane Models
    - Hue
    - Saturation
    - Value
    - HSV
    - Luma
- Image Save etc

Some of these features have been discussed below:

1) Image Addition: Image is a simple matrix. Since addition can be performed on matrix, so that can also be applied on images. To do this a function available in MATLAB is used which performs addition of pixel values of first image to the second image. Addition of the image by itself gives a better clarity of the original image. Fig. 3 demonstrates the image addition operation.

2) Image Clarity: Generally for certain images the pixel distribution is not equally spaced or rather are clotted to a particular intensity level, hence, making the image too dull or too bright [11]. For this reason the following histogram adjustment techniques are being used: Image Adjust, Histogram equalization, Adaptive Histogram equalization. Fig 5. Illustrates the histogram equalization feature.

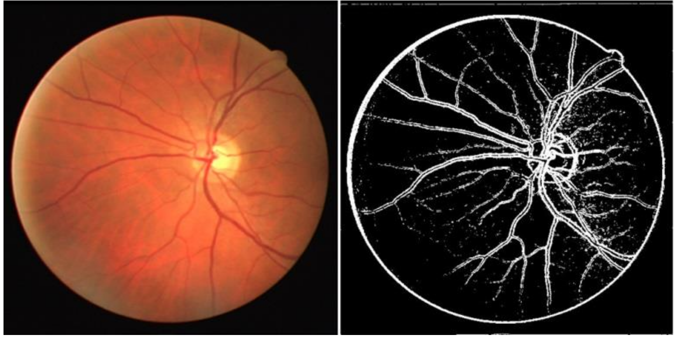


Fig. 2. Extracted Blood Vessels

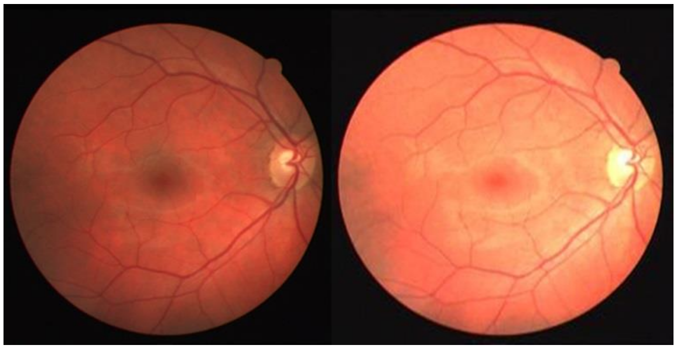


Fig. 3. Demonstration of Addition operation on the image

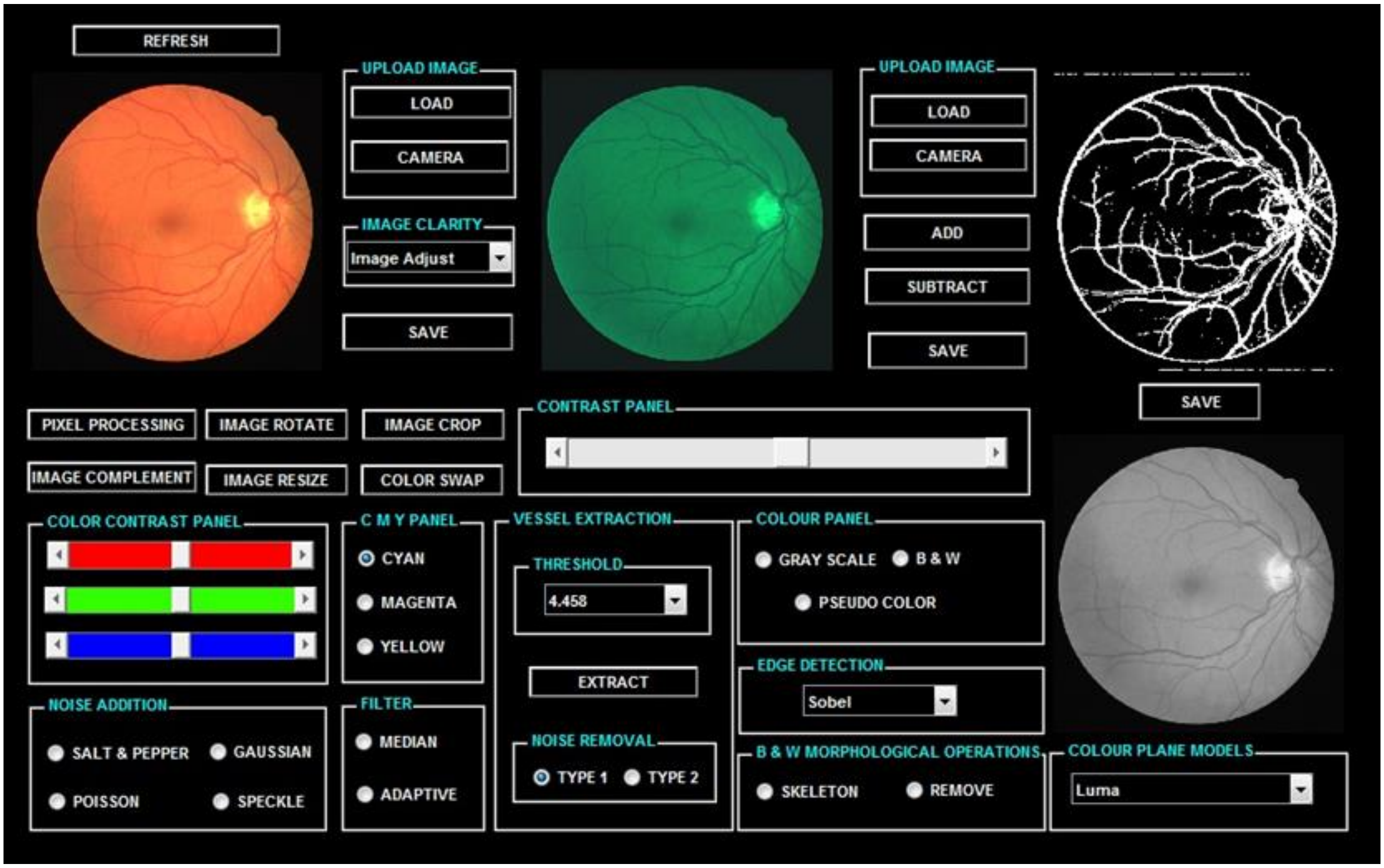


IMAGE CLARITY: Image Adjust
THRESHOLD: 4.458
EDGE DETECTION: Sobel
COLOUR PLANE MODELS: Luma

Fig. 4. Complete MATLAB based GUI demonstrating the extraction of Retinal Blood Vessels Using Kirsch's Templates

3) Image Rotate: Image rotation is used to rotate the image to a specified degree and get a different perspective of view.

4) Image Complement: Image complement simply complements the image. Here, the RGB image of the fundus is complemented where each pixel value is subtracted from the maximum pixel value supported by the class and the difference is used as the pixel value in the output image.

5) Image resize: It is being used to resize the actual image to certain multiples.

6) Image crop: It is used to select any particular portion of the whole image.

7) Contrast Panel: This panel is used to apply brightness or contrast to the image. It is varied over a range of 0 to 255.

8) Colour Contrast Panel: Slider for Red, Green and Blue are added to view the single coloured image in varied brightness (dark and light). It has been mentioned that a colour or RGB image is an overlap of three two dimensional matrices.

9) CMY Panel: This panel is used to view the Cyan, Magenta and Yellow components of the image separately.

10)Colour Panel: The most important feature of this panel is the 'GRAY' radio button that converts an RGB image into a Gray scale image.

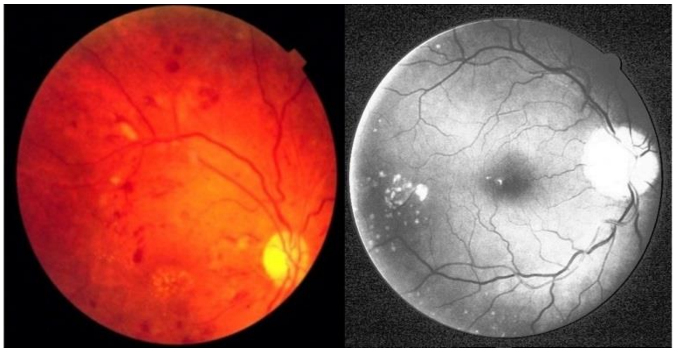


Fig. 5. Image Adjust and Histogram Equalization

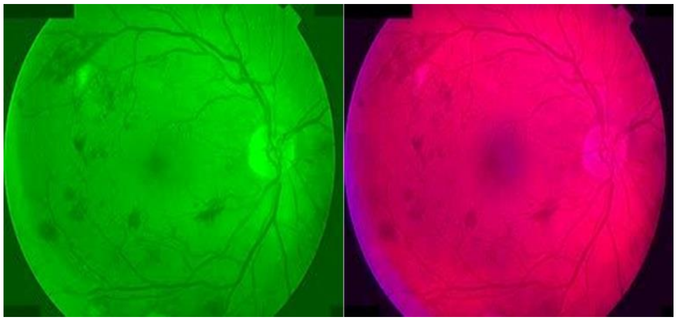


Fig. 6. Green & Magenta Images in addition with contrast

11)Noise addition and removal: Various types of noises get added to an image when a snapshot is taken. In order to get rid of these noises various types of filters are used. To illustrate this we have added a noise to an image externally and then applied various filters to get rid of it and evaluated the results. The types of noises available in this GUI are Poisson, Salt & Pepper, Gaussian and Speckle. Noise removal maybe performed using either a Median filter or an Adaptive filter. Fig. 9. Show the addition and removal of 'salt and pepper noise'.

12)Edge detection: The various types of edge detection techniques are: Sobel, Prewitt, Roberts, LoG, Canny and Zerocross. These options are available for the user to compare the results with those of the Kirsch's Algorithm.

13)Colour Plane Models: In this panel the ophthalmologist would get the opportunity to select any of the following image type which would enable a different cylindrical-coordinate representation of the RGB Colour image:

a) Hue image: Using this option might enable the ophthalmologist to detect further abnormalities of the fundus [12] that are not very evident from the Retinal Blood Vessel Extraction.

b) Saturation image: This option would let the ophthalmologist view the picture of the fundus in a different colour tone as shown in Fig. 8.

c) Hue Saturation Value (HSV) image: The combination of Hue Saturation and Value gives a better outlook of the fundus for detection of abnormalities.

d) Luminosity (luma) image: By this option the ophthalmologist gets the view of the weighted sum of the gamma corrected image of the main RGB image of the fundus.

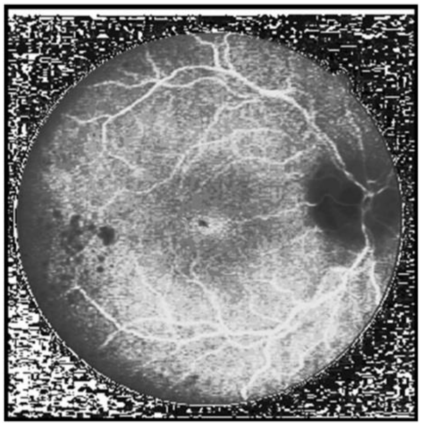


Fig. 7. Hue Image

14)Save: The Save option is provided against the three working axes present in the middle, top right corner and bottom right corner of the GUI for the purpose of saving the edited or worked on images for future reference.

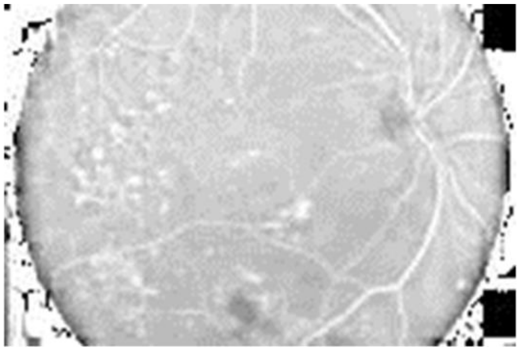


Fig. 8. Saturation Image

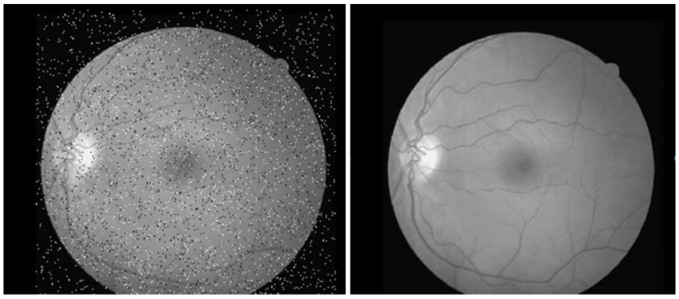


Fig. 9. Demonstration of Addition and Removal of 'Salt & Pepper' Noise using a Median Filter

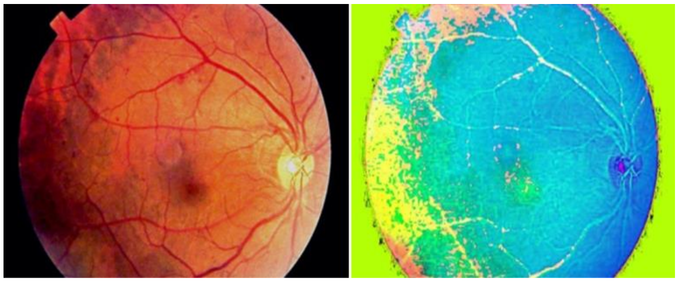


Fig. 10. Demonstration of HSV Image in comparison to the original Fundoscopic Image

IV. Discussions

Diabetic Retinopathy is an important cause of blindness. It is an outcome of prolonged accumulated damage caused to the retinal blood vessels. One percent of global blindness is as a result of Diabetic Retinopathy [13]. We have presented in this paper the retinal blood vessel extraction with the help of *Kirsch's Templates* and a combination of various MATLAB image processing techniques. The combination of these two techniques provided an optimum result. Table1 depicts the comparison between our Results and other Techniques in the extraction of Blood Vessels.

The interface is presented in a user-friendly manner. Individual Save buttons next to each and every image window helps the user to re-use and re-apply the extraction process on edited images. This process enhances the base image at every stage.

Our main aim is to provide a cost effective solution to detect Retinal anomalies and provide a better platform. In some cases, minute examination of Diabetic Retinopathy is carried on and analyzed by Fundus Photography and for detailed insight Optical Coherence Tomography is used [17]. Table 2 represents a comparative study between our Graphical User Interface and the present clinical technique for the detection of Retinal diseases.

TABLE I. COMPARISON OF OUR RESULTS WITH OTHER TECHNIQUES FOR BLOOD VESSEL EXTRACTION

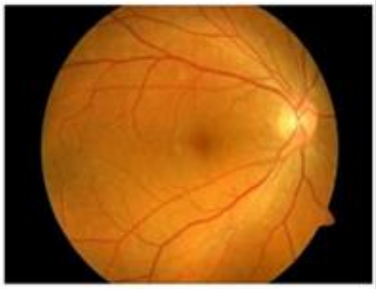
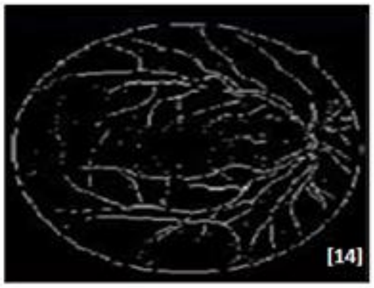
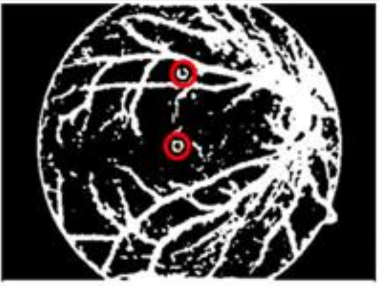
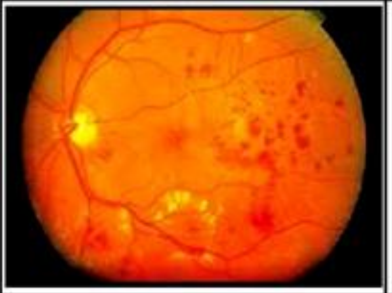
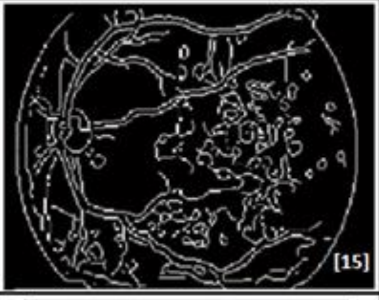
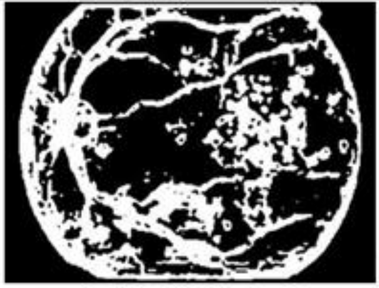
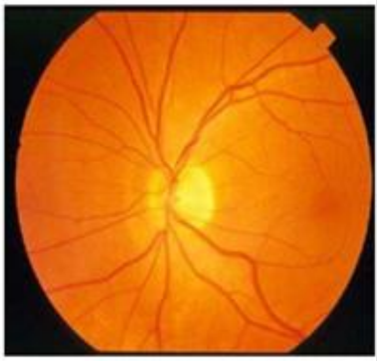
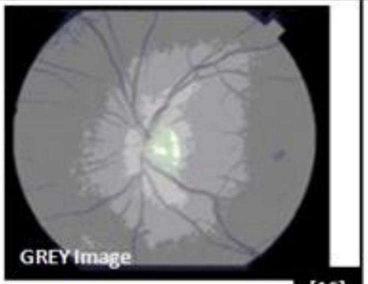
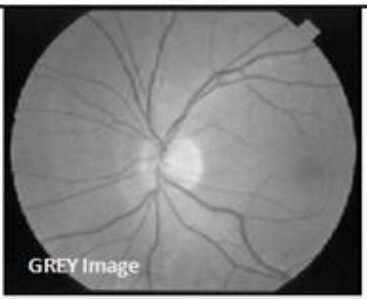
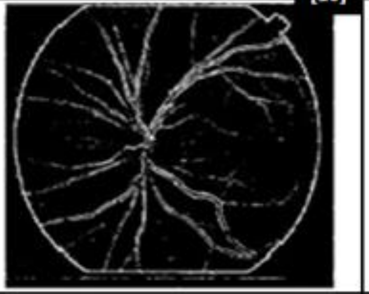
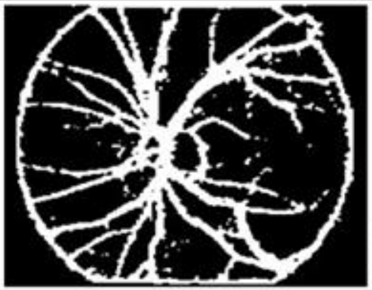
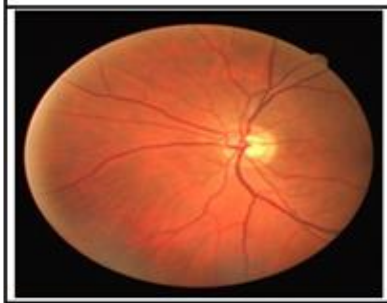
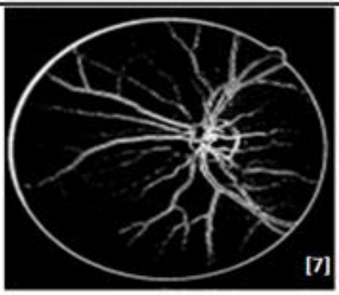
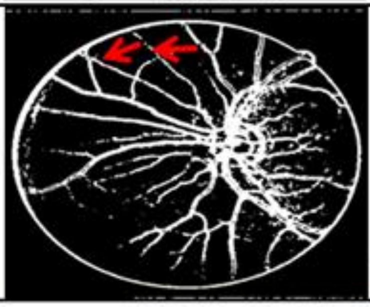
Raw Photo	Compared Algorithm Output	Our Algorithm Output	Discussion
			The ones marked in Red depicts the Lesions.
			The lesions and blood clots are non differentiable.
	 [16]		More Prominent Grey Image.
			As in, the Extracted Blood Vessel depicts more continuity and prominence.
			Blood Vessel continuity is maintained.

TABLE II. COMPARISON OF OUR GRAPHICAL USER INTERFACE WITH PRESENT METHODOLOGY FOR DIAGNOSIS OF RETINAL ABNORMALITIES

Parameters	**Fundus Photography + our Graphical User Interface**	**Fundus Photography + Optical Coherence Tomography (OCT)**
Target Audience	Anyone (A super specialized Doctor or even a Normal Person)	Doctor or Experienced Technician
Ease of Use	No Expertise required	Expertise required
Diagnosis Time	5-10 seconds	Few minutes
Initial Setup	A Computer/Laptop	The OCT machine
Initial Setup Cost	300 USD (approx.)	8995 USD (approx.)
Cost	NIL	Diagnosis Charge

As per Dr.Somnath Das [18], Associate Professor of *Regional Institute of Ophthalmology (RIO), Medical College (Kolkata):*-"This automated system for the extraction of Retinal blood vessels can give us important clue regarding alteration in morphological pattern and pathological changes in and around the retinal blood vessels, status of laminar blood flow within the blood vessels and nature of extravasations of plasma, blood cells and lipids in surrounding retinal tissue on the basis of which we can diagnose and plan for the management of a large group of ocular diseases. Not only that, this system will also be a good prognostic indicator for a particular disease".

Availability of a Computer/Laptop will enable the user to use the Graphical User Interface after a quick installation of MATLAB. One of the limitations of this Graphical User Interface is that it requires a high processor speed for smoothness and trouble-free output. Other than that, slight lag time can be experienced during the execution of the program.

V. FUTURE PROSPECTS

As an Initial stage of our work we have chosen Diabetic Retinopathy as our base to validate the Graphical User Interface performance. We can also utilize our Graphical User Interface to understand pathogenesis and ongoing changes in the retinal blood vessels in diseases like Diabetic Retinopathy, Central Retinal Vein Occlusion(CRVO),Branch Retinal Vein Occlusion(BRVO),Central Retinal Artery Obstruction(CRAO),Hypertensive Retinopathy by which we can make staging of the disease, can assess the severity or actual burden of the disease and can make a suitable management protocol to treat it (LASER photocoagulation, intra vitreal injection of anti-Vascular endothelial growth factor etc.). We can also detect and evaluate diseases like HIV Retinopathy, Arterio-venous malfunctions, Optic Atrophy, Ocular diseases like Ocular Ischaemic syndrome, Arterio-venous malformation in carotid-cavernous fistula and haemangioma, Torch infections having an appearance of CMV (Cytomegalovirus) Retinitis, Salt & Pepper Retinopathy, Rubella etc with the help of our Graphical User Interface. Knowledge about vascular stasis and detection of abnormal optic nerve head perfusion pressure can be extremely helpful as far as the diagnosis of Open angle Glaucoma, Normal tension glaucoma and retinal Vasculitis are concerned. Comparison of structural pattern of retinal vascular tree in Eale's disease, Pan uveitis, Posterior Uveitis and Takayasu's disease involving retina can be done with this method. Furthermore we can incorporate an algorithm to differentiate between blood clots and lesions and provide a comparative study. This would help in distinguishing various Retinal abnormalities. But for all this we require the fundus photography which is obtained by a fundus Camera which costs almost 3600 USD. To cut down the cost at that level, the next step could be utilizing the Smart Phones' Camera and combining it with an extra lens to capture Fundus images more easily and in a very low cost. This whole setup can then be incorporated in Smart phones as an Application which will be easily accessible and utilized by all.

Further classification of special features detected from the extracted blood vessels along with the use of a trained

Probabilistic Neural Network to recognise and report any of the above mentioned abnormalities can be carried on.

Further classification of special features and parameters detected from the extracted blood vessels along with the use of a trained Probabilistic Neural Network to recognize and report any of the above mentioned abnormalities can be carried out automatically. This Automation of disease detection may also be achieved by the implementation of Neuro-fuzzy template, since it provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing information from the parameters of the pre-processed fundus image for each disease.[19].

VI. CONCLUSIONS

Retinal images are being used by ophthalmologists to aid in diagnosis, to make measurements, and to look for changes in lesions or severity of diseases. The appearance of the retinal vasculature particularly acts as an indicator for diagnosis of Diabetic Retinopathy (both Proliferative and Non-Proliferative) and Glaucoma. Therefore, extraction of these features is the key challenge for proper analysis, visualization and quantitative comparison. The present study focuses mainly on this challenge of blood vessel extraction from colour retinal images obtained from fundoscopy. The proposed algorithm proves to be successful and robust in accurately extracting the retinal blood vessels. In this respect, the dataset of 40 test images from the DRIVE database has been used to evaluate this method. Few real time images obtained from the Regional Institute of Ophthalmology (RIO), Medical College, Kolkata are also used for the evaluation. It included fundoscopy of various patients having NPDR (Non

Proliferative Diabetic Retinopathy), PDR (Proliferative Diabetic Retinopathy) and post PRP (Panretinal Photocoagulation). The fact that the proposed algorithm extracts blood vessels with evident accuracy renders it a quite sought after platform for further improvements. An accurate extraction of blood vessels provides the basis for the measurements of a variety of features including micro aneurysms and hard exudates that can then be applied to the tasks of diagnosis, treatment, evaluation and clinical study [20]. More particularly in rural areas of underdeveloped and developing countries the proposed Graphical User Interface can be utilized to overcome the hardships arising due to the shortage of professional observers by this completely automated computer assisted monitoring and diagnostic system.

ACKNOWLEDGEMENT

This work was supported by ESL (www.eschoollearning.net), Kolkata.

REFERENCES

[1] Muthukrishnan, R., Radha, M, (2011), "Edge Detection Techniques for Image Segmentation", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No.6, pp. 259-267,DOI:10.5121/ijcsit.2011.3620.

[2] Yin,Y. , Adel, M, Bourennane, S., (2012), "Retinal vessel Segmentation using a Probabilistic Tracking Method", *Pattern Recognition*, Vol. 45, No. 4, pp. 12335-1244, DOI: 10./1016/j.patcog.2011.09.019.

[3] Zhang, B., Zhang, L., Zhang, L., Karray, F. (2010), "Retinal vessel extraction by matched filter with first-order derivative of Gaussian", *Computers in Biology and Medicine, Vol*. 40, No. 5, pp. 438-445,DOI:10.1016/j.compbiomed.2010.02.008.

[4] Esmaeili, M., Rabbani, H., Mehri, A., Dehghani, A., (2009), "Extraction of retinal blood vessels by curvelet transform", 16th IEEE International Conference on Image Processing (ICIP), pp.33533356, DOI10.1109/ICIP.2009.5413909.

[5] Niemeijer, M., Staal, J., van Ginneken, B., Loog, M., and Abramoff, M., "Comparative study of retinal vessel segmentation methods on a new publicly available database," *SPIE Medical Imaging*, vol. 5370, pp. 648– 656, 2004

[6] *Regional Institute of Ophthalmology (RIO), Medical College (Kolkata).*

[7] Karasulu,"Automatic Extraction of Retinal Blood Vessels: A Software Implementation", European Scientific Journal December edition vol.8, No.30 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431.

[8] Gao, P., Sun, X., Wang, W., (2010), "Moving Object Detection Based on Kirsch Operator Combined with Optical Flow", International Conference on Image Analysis and Signal Processing (IASP), 9-11 April 2010, pp. 620-624. DOI: 10.1109/IASP.2010.5476045.

[9] Vijayakumari,V. Suriyanarayanan, N., "Survey on the Detection Methods of Blood Vessels in retinal Images", European journal of scientific research ISSN 1450-216X Vol. 68 No.1(2012).

[10] Cemil Kirbas and Francis Quek, "A Review of Vessel Extraction Techniques and Algorithms", ACM computing surveys, Vol. 36, No.2, June 2004.

[11] Raquib Buksh, Soumyajit Routh, Parthib Mitra, Subhajit Banik , Abhishek Mallik, Sauvik Das Gupta "MATLAB based Image Editing and Color Detection", International Journal of Scientific and Research Publications,

[12] Hayashi, J., Kunieda, T., Cole, J., Soga, R., Hatanaka, Y., Lu,M., Hara, T., and Fujita, F., "A development of computer-aided diagnosis system using fundus images", Proc. Of IC on (VSMM 2001), pp. 429-438, 2001.

[13] Global data on visual impairments 2010. Geneva, World Health Organization, 2012.

[14] M.Kavitha, Dr.S.Palani, "A New Fast Curvelet Transform with Morphological Operations based method for Extraction of Retinal blood vessels using Graphical User Interface", International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012, ISSN 2229-5518 .

[15] Archna Sharma and Hempriya, "Detection of Blood Vessels and Diseases in Human Retinal Images", ISSN 2319-7080, International Journal of Computer Science and Communication Engineering IJCSCE Special issue on "Emerging Trends in Engineering & Management" ICETE 2013.

[16] H.S. Bhadauria1, S.S. Bisht and Annapurna Singh, "Vessels Extraction from Retinal Images", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 2278-8735. Volume 6, Issue 3 (May. - Jun. 2013), PP 79-82.

[17] David J Browning, , Michael D McOwen, Robert M Bowen, Tisha L O'Marah, "Comparison of the clinical diagnosis of diabetic macular edema with diagnosis by optical coherence tomography", Presented at: American Academy of Ophthalmology Annual Meeting, November, 2003; Anaheim, California.

[18] Dr. Somnath Das, Associate Professor, Regional Institute of Ophthalmology (RIO), Medical College (Kolkata).

[19] Bhijupukan Bhagapathi ChumiDas, "Edge Detection of Digital ImagesUsing Fuzzy Rule Based Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.)

[20] Osareh, B. Shadgar, "Automatic Blood Vessel Segmentation In Color Images Of Retina", Iranian Journal of Science & Technology, Transaction B,Engineering, Vol. 33, No. B2, pp 191-206, 2009.

An Artificial Neural Network Application for Estimation of Natural Frequencies of Beams

Mehmet Avcar
Department of Civil Engineering
Faculty of Engineering, Suleyman Demirel University
Isparta, Turkey

Kemal Saplıoğlu
Department of Civil Engineering
Faculty of Engineering, Suleyman Demirel University
Isparta, Turkey

***Abstract*—In this study, natural frequencies of the prismatical steel beams with various geometrical characteristics under the four different boundary conditions are determined using Artificial Neural Network (ANN) technique. In that way, an alternative efficient method is aimed to develop for the solution of the present problem, which provides avoiding loss of time for computing some necessary parameters. In this context, initially, first ten frequency parameters of the beam are found, where Bernoulli-Euler beam theory was adopted, and then natural frequencies are computed theoretically. With the aid of theoretically obtained results, the data sets are formed and ANN models are constructed. Here, 36 models are developed using primary 3 models. The results are found from these models by changing the number and properties of the neurons and input data. The handiness of the present models is examined by comparing the results of these models with theoretically obtained results. The effects of the number of neurons, input data and training function on the models are investigated. In addition, multiple regression models are developed with the data, and adjusted R-square is examined for determining the inefficient input parameters**

Keywords*—natural frequency; beam; ANN; multiple regression; adjusted R-square

I. INTRODUCTION

Every structure in the nature has endless number of vibration frequencies and mode shapes, and calculation of these frequencies and their mode shapes are important to solve the vibration induced engineering problems [1-5].

Vibration analyses of structural systems have been performed with the aid of different methods [6-15]. However, the complex shaped structures may be analyzed with soft computing techniques more easily. Soft Computing is a general term for a collection of computing techniques [16]. These well-known techniques constitute artificial neural networks (ANN), fuzzy logic, evolutionary computation, machine learning and probabilistic reasoning. Soft computing methods differ from classical computing methods in that, unlike classical computing methods it is tolerant of imprecision, uncertainty, partial truth to achieve tractability, approximation, robustness, lows solution cost and better rapport with reality [17].

Although all above mentioned techniques have been adapted to the structural analysis, design and optimization problems, especially ANNs have been widely used in many fields of science and technology, such as, in vibration problems of engineering structures, due to it has an excellent learning capacity [18]. Gates et al. [19] presented a method of using artificial neural networks stabilizing large flexible space structures, in which the neural controller learns the dynamics of the structure to be controlled and constructs control signal stabilizing structural vibrations. Karlik et al. [20] studied the nonlinear vibrations of an Euler-Bernoulli beam with a concentrated mass using ANN technique which has a multi-layer, feed-forward, back propagation algorithm. Mahmoud and Kiefa [21] investigated the feasibility of using general regression neural networks to solve the inverse vibration problem of cracked structures, in which a steel cantilever beam with a single edge crack is examined as a case study. Castillo et al. [22] presented a general methodology to develop and work with functional networks, which is a network based alternative to the neural network paradigm. Cevik et al. [23] suggested ANN approach for obtaining the natural frequencies of suspension bridges. Civalek [24] examined flexural and axial vibration of elastic beams with various support conditions using ANN, in which the first three natural frequencies of beams are obtained using multi-layer neural network based back-propagation error learning algorithm. Hassanpour et al. [25] investigated the vibration of the simply-supported beam with rotary springs at either ends using a multilayer feed-forward back-propagation ANN. Bağdatlı et al. [26] studied the nonlinear vibrations of stepped beam systems using ANN technique which has a multi-layer, feed-forward, back-propagation algorithm networks. Saeed et al. [27] presented various artificial intelligence techniques for crack identification in curvilinear beams based on changes in vibration characteristics. Jalil et al. [28] presented dynamic model of flexible cantilever beam in transverse motion using finite difference approach, in which the identification of a flexible beam structure was utilized using neural network. Mohammadhassani et al. [29] presented comparison of the effectiveness of artificial neural network and linear regression in the prediction of strain in tie section using experimental data from eight high-strength-self-compact concrete deep beams.Ding et al. [30] determined locating and quantifying damage in beam-type structures using structural dynamics-guided hierarchical neural-networks scheme. Karimi et al. [31] suggested an alternative modeling technique using ANN for predicting the effects of different parameters on the natural and nonlinear frequencies of the laminated plates.

In the present study bending natural frequencies of the prismatical steel beams with various geometrical characteristics under the four different boundary conditions, i.e. Clamped-Clamped (C-C), Clamped-Free (C-F), Clamped-Simply

Supported (C-SS) and Simply Supported-Simply Supported (SS-SS) is determined using ANN technique. Initially, the first ten natural frequency parameters of the beam are found adopting Bernoulli-Euler beam theory, and then natural frequencies are computed theoretically. With the aid of theoretically obtained results the data sets are formed and ANN models are constructed. Here, 36 models are developed using primary 3 models. The results are found from these models by changing the number and properties of the neurons and input data. The handiness of the present models is examined by comparing the results of these models with theoretically obtained results. The effects of the number of neurons, input data and training function on the models are investigated. In addition, multiple regression models are developed with the data, and adjusted R^2 is investigated for determining the inefficient input parameters. To the best of authors knowledge, although various studies are presented on the free vibration analysis of the structures using ANN technique, the effects of the number of neurons, input data and training function on the models are not investigated in detail, and the inefficient input parameters are not determined using multiple regression models. In the present work, an attempt is made for addressing these issues.

II. Mathematical Modelling of the Problem

Consider an elastic beam of length L, width b, height h, Young's modulus E, and mass density ρ with uniform cross section A, as shown in Fig. 1.

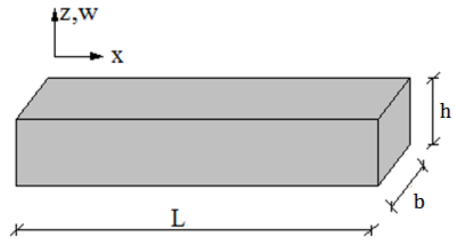


Fig. 1. Geometry of the beam

Using Euler-Bernoulli beam theory, one can obtain the equation of motion of a beam with homogeneous material properties and constant cross section as follows [1-5]

$$\frac{\partial^4 w}{\partial x^4} + \mu\frac{\partial^2 w}{\partial t^2} = 0 \tag{1}$$

where the following definition apply

$$\mu = \frac{\rho A}{EI} \tag{2}$$

here I is the area moment of inertia of the beam cross section, w is the transverse displacement, and t is time.

The solution of the Eq. (2) is sought by separation of variables. Therefore, the displacement is separated into two parts: one is depending on the position and the other is depending on time:

$$w(x,t) = \alpha(x)\beta(t) \tag{3}$$

where α and β are independent of time and position, respectively.

Substituting Eq. (3) into Eq. (2) and after some mathematical rearrangements, the following equation is obtained:

$$-\frac{1}{\mu\alpha(x)}\frac{\partial^4\alpha(x)}{\partial x^4} = \frac{1}{\beta(t)}\frac{\partial^2\beta(t)}{\partial t^2} = -\omega^2 \tag{4}$$

Here the each side resulting equation is set to equal a constant, denoted $-\omega^2$, to have simple harmonic motion in the beam.

If the position variable of Eq. (4) is separated

$$\frac{\partial^4\alpha(x)}{\partial x^4} - \lambda^4\alpha(x) = 0 \tag{5}$$

where

$$\lambda^4 = \omega^2\mu \tag{6}$$

If the time variable is separated

$$\frac{\partial^2\beta(t)}{\partial t^2} + \omega^2\beta(t) = 0 \tag{7}$$

Eq. (5) is solved as follows:

$$\alpha(x) = A_1\sinh\lambda x + A_2\cosh\lambda x + A_3\sin\lambda x + A_4\cos\lambda x \tag{8}$$

where A_1, A_2, A_3, A_4 are constants, $\sinh$ and $\cosh$ are the hyperbolic $\sin e$ and $\cos e$ functions, respectively.

Eq. (7) is solved as follows:

$$\beta(t) = A_5\sin\omega t + A_6\cos\omega t \tag{9}$$

where A_5 and A_6 are constants.

Thus, if Eq. (8) is multiplied by Eq. (9) to obtain $w(x,t)$, it yields eight combined constants as:

$$w(x,t) = \left(A_1\sinh\lambda x + A_2\cosh\lambda x + A_3\sin\lambda x + A_4\cos\lambda x\right) \\ \times\left(A_5\sin\omega t + A_6\cos\omega t\right) \tag{10}$$

where the constants A_1, A_2, A_3, A_4 can be obtained from the boundary conditions, and A_5, A_6 can be obtained from the initial conditions

The boundary conditions satisfied by a C-C, C-F, C-SS, SS-SS beams are as follows, respectively:

$$w(0) = w'(0) = w(L) = w'(L) = 0 \tag{11}$$

$$w(0) = w'(0) = w''(L) = w'''(L) = 0 \tag{12}$$

$$w(0) = w'(0) = w(L) = w''(L) = 0 \tag{13}$$

$$w(0) = w''(0) = w(L) = w''(L) = 0 \tag{14}$$

Substituting boundary conditions given in Eqs (11-14) into Eq. (8) separately; and then after some mathematical operations, the frequency parameters of the beam, $\eta = \lambda L$, are obtained for the first ten modes. Finally, using Eq. (6) the natural frequency f_n (Hz) of the beam is found as follows:

$$f_n = \frac{\omega}{2\pi} \qquad (15)$$

III. ANN Modelling of the Problem

A. Structure of ANN

ANN is a technique that seeks to build an intelligent program using models that simulate the working network of the neurons in the human brain (Fig. 2). Unlike conventional computational programs, the ANN does not have exact data and provides outputs with respect to introduced data set. The data and the circumstances introduced to the program are put into process by the help of various methods of education and learnings. With the aid of the outputs of these transactions, the program assigns weights between the data and the neurotic structures. Afterward, when come up to different situations and data, the cases are commented and results are presented in accordance with previous learnings [32].

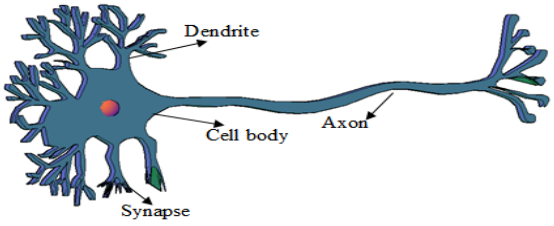


Fig. 2. A biologic nerve cell structure

The basic unit of ANN is called as a process element or a node. Although the artificial nerve elements are simpler than the biological nerves, it can simulate the 4 main functions of biological nerves (Fig. 3).

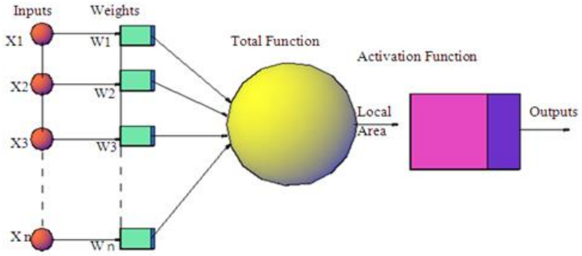


Fig. 3. Artificial neural network sample

There are plenty of neural network models in the existing literature. However, the most preferred neural network model is back propagation model. It is experienced that this model gives pretty good results in the estimation and classification processes [33]. Back propagation neural network is the mostly preferred model because of its capability and excellence to solve problems which are nonlinear and have very complicated structures. Back propagation neural network is a multi-layer and feed-forward neural network trained by the Back Propagation algorithms [7]. This model makes weight assignment processing the inputs and the outputs again and again, and the model tries to minimize the least square errors using this operation. The mathematical expression of this model is as follows [34]

$$\Delta W_n = a\Delta W_{n-1} - b_T \frac{\partial F}{\partial W} \qquad (16)$$

Here, w is a value of assigned weight between any two neurons, ΔW_n and ΔW_{n-1} are respectively the changes of weightings for n and n-1 values, a is the coefficient of momentum, b_T is the ratio of training, F is the calculated error

where

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_i - P_i) \qquad (17)$$

here T_i is the actual output or namely target and P_i is the estimated output value. The working principle of the ANN model is shown in Fig. 4, the inputs are included into the model after weight adjustment. The data are forwarded to the activation function being processed in each neural network with the weights. The results are compared with the actual results in order to determine error. The errors found are transferred to initial weights with the help of Back Propagation and this process is repeated for a number of times which is called as Epoch. Once the process is completed, the results with minimum errors are found [35].

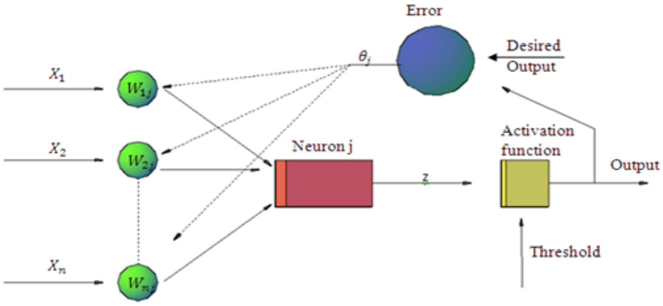


Fig. 4. Neuron weight adjustments

B. Normalizing the data

The input and output values are required to be restricted in some certain rules for artificial neural network models. This process is called as normalization. The most used normalization functions are Min Rule, Max Rule, Median, Sigmoid and Z-Score [32]. In this study, Min-Max normalization rule is applied as below:

$$Z' = \frac{Z_i - Z_{min}}{Z_{max} - Z_{min}} \qquad (18)$$

Here, Z' is the normalized data, Z_i is the actual data, Z_{max} is the maximum data and Z_{min} is the minimum data. Through this equation, all the data are normalized in the range of [0-1]. Hereby, both the error distribution is done in a narrower range and model runs more quickly.

C. Multiple regression analysis

Multiple regression analysis is applied for evaluating the effect of multiple independent variables (x) on a dependent variable (y). In multiple linear regression analysis, it is assumed that each independent variable has a relationship with the dependent variable [36]. This relationship is expressed as below:

$$y = c + b_1x_1 + b_2x_2 + ... + b_nx_n \qquad (19)$$

Here, c is a constant number, b_i are the coefficient of the variables.

To calculate the coefficients in the Eq. (19), the mean square method is used. The difference between the actual y and the theoretical y is minimized as follow

$$\sum_{i=1}^{n} y_i - \left(c + B_1x_{1i} + B_2x_{2i} + ... + B_nx_{ni}\right) \qquad (20)$$

In order to evaluate the accuracy of multiple regression model, the regression coefficient is required to be determined. Besides, multiple regression is applied with respect to Stepwise Selection Method for determining the necessity of the parameters.

IV. Results and Discussion

In this section, natural frequencies, $f_n\,(Hz)$, of prismatical steel beams under four different boundary conditions are examined. For this aim, at first natural frequency parameters are obtained, then natural frequencies, $f_n\,(Hz)$, of prismatical steel beams are found theoretically. Afterward, from obtained these results data sets are constructed. Here, a total of 8640 data sets are used in training stage, and a total of 1920 data sets are used in the testing stage. By this way, 3 main models and a total of 36 sub-models created by changing the number of neurons of the main models. The sizes, moment of inertia and boundary conditions are used as input data parameters. The natural frequency values are employed as outputs. Input data parameters and their intervals and mechanical parameters of steel are given in Table 1 and Table 2, respectively.

TABLE I. The Input Data Parameters and Their Intervals

Parameter	Minimum	Maximum
b (m)	0.10	0.15
h (m)	0.10	0.15
L(m)	3	3.5
I (m^4)	8.333×10^{-6}	4.219×10^{-5}
n	1	10
Case	1	4

Here, b, h, L and I denote width, height, length and moment of inertia of the beam, and n denotes mode number, Case 1,2,3,4 denoted in C-C, C-F, C-SS and SS-SS boundary conditions, respectively.

TABLE II. The Mechanical Parameters of Steel

Parameter	Description
Young's Modulus (E)	$2.1\times10^{11}\ N/m^2$
Passion Ratio (ν)	0.3
Density (ρ)	$7850\,kg/m^3$

A. Numerical examples

Example 1: In this example, a comparative study is performed to validate the present numerical results. For this purpose, theoretically obtained exact results of natural frequency of the prismatical beams under the four different boundary conditions versus mode number (n) are compared with those obtained using ANN, in Table 3. It is found that the numerical results of both methods are consistent, which show the accuracy of the present ANN model. The absolute errors are calculated as follows: $\left|\frac{f_{nANN} - f_{nExact}}{f_{nExact}}\right| \times 100$. Besides, the variations of absolute errors in the natural frequencies, $f_n\,(Hz)$, are illustrated in Fig. 5.

TABLE III. Variations of Natural Frequencies of the Prismatical Steel Beams Under the Four Different Boundary Conditions Versus Mode Number (N) ($b = 0.14m; h = 0.14m; L = 3m$)

n	f_n(Hz)							
	Case 1		**Case 2**		**Case 3**		**Case 4**	
	Exact	***ANN***	***Exact***	***ANN***	***Exact***	***ANN***	***Exact***	***ANN***
1	80.71	81.11	12.68	12.27	55.62	56.14	35.60	35.65
2	222.47	224.35	79.48	77.56	180.24	183.16	142.41	140.96
3	436.14	436.24	222.47	222.18	376.06	377.06	320.43	318.73
4	720.97	719.61	436.14	433.72	643.09	642.07	569.65	567.22
5	1077.01	1075.64	720.97	718.09	981.33	979.88	890.09	887.23
6	1504.26	1503.92	1077.01	1076.40	1390.77	1390.24	1281.73	1282.26
7	2002.71	2003.94	1504.26	1503.87	1871.42	1872.75	1744.58	1743.41
8	2572.37	2572.38	2002.71	2000.87	2423.28	2424.99	2278.64	2278.86
9	3213.23	3209.13	2572.37	2573.64	3046.34	3041.29	2883.90	2884.54
10	3925.31	3933.66	3213.23	3210.43	3740.61	3749.58	3560.37	3557.60

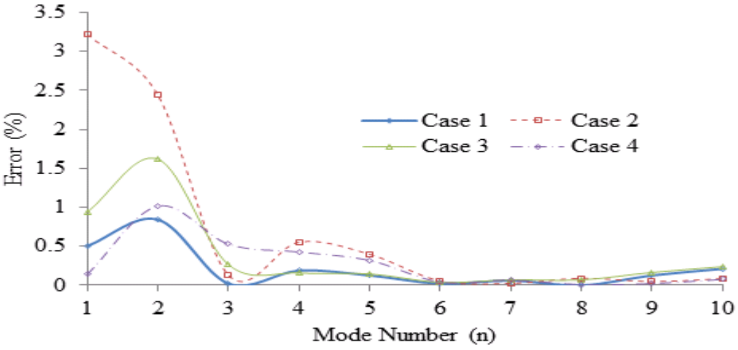


Fig. 5. Variations of error in the natural frequencies of the prismatical steel beams under the four different boundary conditions versus mode number (n)

Example 2: Table 4 shows the Model 1, in which 6 inputs, 1 hidden layer and 1 output (see Fig. 6) are used for obtaining the natural frequencies, of the prismatical steel beams under the four different boundary conditions.

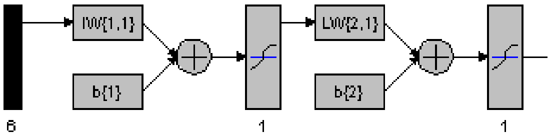


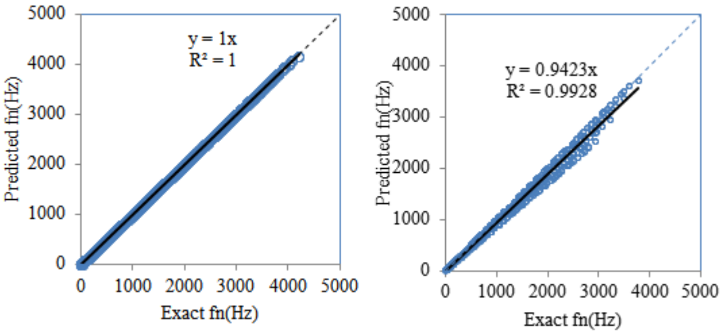
Fig. 6. Schematic architecture of Model 1 (6-1-1)

As with all models, feed-forward back propagation algorithm is used in the network type. The tangent and logarithmic sigmoid transfer functions are employed. The number of hidden layer is determined as 1, and 12 separate sub-models are created using the models having 1 to 9 neurons. Actual output values of the natural frequencies are compared with those obtained from training. The results are found with the very small errors, especially for the models with 5 and more neurons.

TABLE IV. TRAINING AND TEST RESULTS FOR MODEL 1

Transfer Function	Number of Neurons	Training Data			Test Data		
		R²	*Equation sets*	*MSE %*	*R²*	*Equation sets*	*MSE %*
Tan Sig.	1	0.996	y=0.9961x	25.66	0.9913	y=0.9364x	31.23
	3	0.9998	y=0.9998x	6.18	0.9933	y=0.938x	12.05
	5	1	y=0.9999x	1.80	0.9932	y=0.946x	6.81
	7	1	y=0.9999x	1.75	0.9932	y=0.9442x	7.16
	8	1	y=x	0.71	0.9931	y=0.9428x	6.71
	9	1	y=0.9999x	1.67	0.9929	y=0.9423x	8.04
Log Sig.	1	0.9959	y=0.9964x	25.66	0.9913	y=0.9364x	31.23
	3	0.9997	y=0.9997x	8.01	0.9933	y=0.938x	11.40
	5	0.9999	y=0.9999x	2.53	0.9932	y=0.946x	7.81
	7	0.9999	y=0.9999x	1.68	0.9932	y=0.9442x	7.74
	8	0.9999	y=x	0.83	0.9929	y=0.9419x	6.88
	9	0.9999	y=0.9999x	1.74	0.9929	y=0.9423x	7.40

To eliminate the possibility of rote learning of these results, 1920 data sets, which are allocated for test, are also included into the model and the obtained results are compared with the exact values. The test data shows that, the interval of error is nearly 6-8%, and regression coefficient takes values very close to 1 for the models with 5 and more neurons. In addition, the best agreement is observed in the model with 8 neurons and plotted in Fig. 7.



a) b)

Fig. 7. Scatter diagrams of Model 1 for tangent sigmoid transfer function with 8 neurons a) training b) test

Example 3: Table 5 shows the Model 2, in which 5 inputs, 1 hidden layer and 1 output are used for obtaining the natural frequencies of the prismatical steel beams under the four different boundary conditions. In this model, moments of inertia, (I), is removed from input parameters and a model with 5 input is created (5-1-1). In the training process of the model for all models of 5 neurons and higher, the error seems to fall below 1%.

TABLE V. TRAINING AND TEST RESULTS FOR MODEL 2

Transfer Function	Number of Neurons	Training Data			Test Data		
		R²	***Equation sets***	***MSE %***	***R²***	***Equation sets***	***MSE %***
Tan Sig.	1	0.9959	y=0.9961x	12.93	0.9912	y=0.9393x	31.22
	3	0.9998	y=0.9997x	1.81	0.9932	y=0.9424x	11.67
	5	0.9998	y=0.9999x	0.87	0.9930	y=0.9425x	8.11
	7	0.9999	y=0.9999x	0.35	0.9937	y=0.9435x	6.80
	8	0.9999	y=0.9999x	0.43	0.9930	y=0.9428x	6.83
	9	0.9999	y=0.9999x	0.53	0.9912	y=0.9425x	7.34
Log Sig.	1	0.9958	y=0.9960x	12.93	0.9922	y=0.9398x	29.22
	3	0.9999	y=0.9998x	1.72	0.9935	y=0.9429x	11.56
	5	0.9999	y=0.9999x	1.04	0.9936	y=0.9437x	8.95
	7	0.9999	y=0.9999x	0.61	0.9940	y=0.9437x	6.65
	8	0.9999	y=0.9999x	0.61	0.9943	y=0.9442x	6.59
	9	0.9999	y=0.9999x	0.96	0.9935	y=0.9435x	6.79

Considering the errors, the model for logarithmic sigmoid transfer function with 8 neurons is found the best and illustrated in Fig. 8.

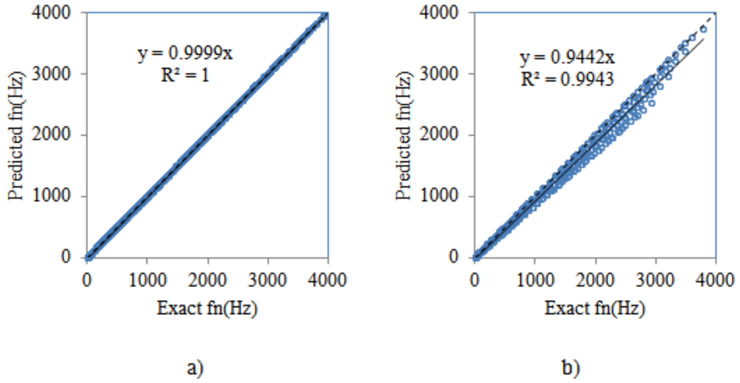


Fig. 8. Scatter diagrams of Model 2 for logarithmic sigmoid transfer function with 8 neurons a) training b) test

It should be noted that, although having one missing input parameter in comparison with the Model 1, the present model does not show serious differences in error rates for both training and test results.

Example 4: Table 6 shows the Model 3. Here parameters are tried to be identified with numbers from 1 to 10 instead of calculations of frequency parameter value. Therefore, the natural frequency parameters are aimed to be determined with ANN model without any calculation in advance. As the model results are examined, it is found that the results are worse than the results of previous ANN models in terms of both training and test. However, in other models, either moment of inertia or natural frequency parameters are included into the model with pre-calculating.

For this model, all the inputs are introduced to the model with simple numeric expressions and the outputs are obtained. As a result, the errors are found very minimal as being 8.93% for the test of the model for logarithmic sigmoid transfer function with 9 neurons and plotted in Fig. 9.

Example 5: Table 7 shows Model 4, in which the training inputs are implanted to the model step by step and adjusted R^2 values are investigated. The step in which the adjusted R^2 values decrease or remain constant; the included parameters are excluded from the model.

According to the steps of the process of Table 7, the input of moment of inertia is excluded and a multiple regression model is constructed with the remaining 5 input parameters in Table 8.

As shown in Fig. 10, the results found in the regression models are remarkably incorrect for both training and test data in comparison with the results obtained from three ANN models. And these results show that the constructed three ANN models give more efficient results for the present problem.

TABLE VI. TRAINING AND TEST RESULTS FOR MODEL 3

Transfer Function	Number of Neurons	Training Data			Test Data		
		R^2	*Equation sets*	*MSE %*	*R^2*	*Equation sets*	*MSE %*
Tan. Sig.	1	0.9784	y=0.9890x	17.72	0.9731	y=0.9321x	38.97
	3	0.9934	y=0.9955x	7.00	0.9867	y=0.9407x	18.75
	5	0.9995	y=0.9992x	2.39	0.9930	y=0.9417x	11.06
	7	0.9998	y=0.9987x	3.24	0.9930	y=0.9418x	12.20
	8	0.9998	y=0.9989x	3.22	0.9930	y=0.9418x	10.31
	9	0.9999	y=0.9999x	1.68	0.9930	y=0.9424x	9.64
Log. Sig.	1	0.9784	y=1.0288x	16.63	0.9731	y=0.9507x	39.18
	3	0.9934	y=1.0189x	7.32	0.9867	y=0.9598x	18.50
	5	0.9995	y=1.0159x	3.43	0.9930	y=0.9602x	10.32
	7	0.9998	y=1.0175x	4.32	0.9930	y=0.9586x	11.47
	8	0.9998	y=1.0179x	4.46	0.9930	y=0.9605x	9.49
	9	0.9999	y=1.0118x	3.06	0.9930	y=0.9612x	8.93

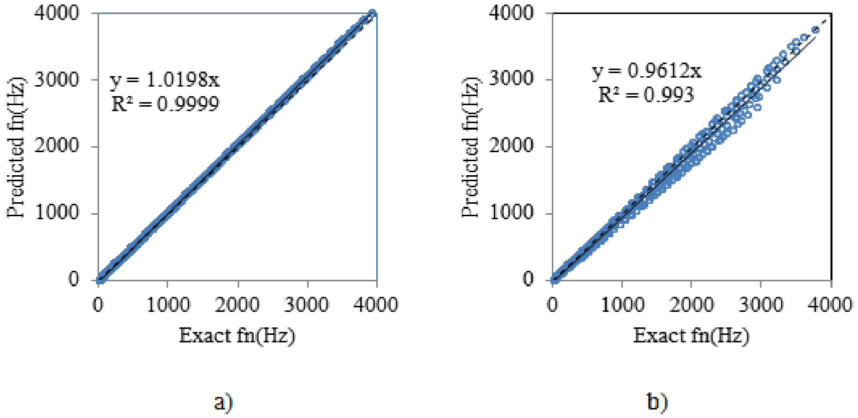


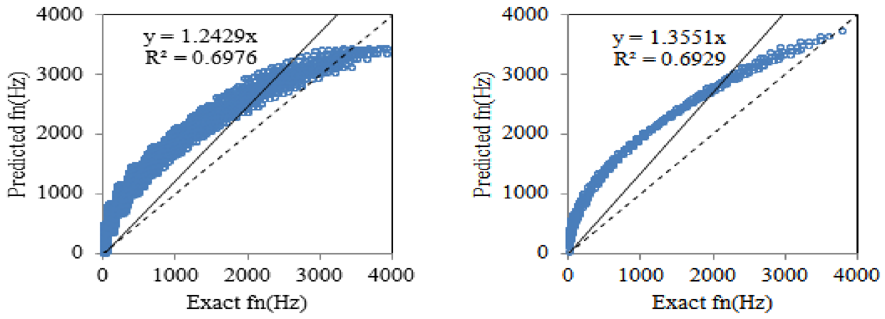
Fig. 9. Scatter diagrams of Model 3 for logarithmic sigmoid transfer function with 9 neurons a) Training b) Test

TABLE VII. ADJUSTED R^2 ANALYSIS RESULTS

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	0.966	0.932	0.932	259.46804

TABLE VIII. MULTIPLE REGRESSION ANALYSIS RESULTS

	Unstandardized Coefficients		Standardized Coefficients	95.0% Confidence Interval for B	
	B	*Std. Error*	*Beta*	*Lower Bound*	*Upper Bound*
Constant	-728.779	106.944		-938.414	-519.144
Case	-2.113	2.498	-0.002	-7.010	2.783
B	0.100	163.450	0.000	-163.350	163.550
H	9238.788	163.450	0.158	8918.388	9559.188
L	-156.070	32.735	-0.013	-220.239	-91.901
η	325.397	.960	0.952	323.517	327.278



a) b)

Fig. 10. Scatter diagrams of multiple regression model a) Training b) Test

V. Conclusion

In this study, bending natural frequencies of the prismatical steel beams with various geometrical characteristics under the four different boundary conditions are determined using ANN technique. In that way, an alternative efficient method is aimed to develop for the solution of the existing problem, which provides avoiding loss of time for computing some necessary parameters.

Briefly the following results are obtained:

1) The tangent sigmoid transfer function shows better performance in Model 1 with 8 neurons

2) The logarithmic sigmoid transfer function shows better performance in Model 2 with 8 neurons

3) When the first two models are considered it is concluded that ANN models do not need moment of inertia parameter in the training

4) The logarithmic sigmoid transfer function provides better results in Model 3 with 9 neurons

5) It is found from Model 4 that the moment of inertia has not any efficiency. This finding also supports the results found by the Models 1 and 2

6) The errors of the models having 5 or more neurons are lower, and this prove at least 5 neurons should be used for the reliably of the model

7) The two transfer functions have quite similar errors, and so both of them can be used

8) The constructed ANN models give more efficient results than the multiple regression model

References

[1] S. P. Timoshenko, Vibration Problems in Engineering, D. Van Nostrand, Princeton, NJ, 1937.

[2] E. B. Magrab, Vibration of Elastic Structural Members, Netherlands, Sijthoff and Noordhoff, 1979.

[3] A. W. Leissa and M. S. Qatu Vibration of Continuous Systems, McGraw Hill Companies, 2011.

[4] S. S. Rao, Vibration of Continuous Systems, John Wiley & Sons, Inc., Hooken, New Jersey, 2007.

[5] C. Y. Wang and C. M. Wang, Structural Vibration: Exact Solutions for Strings, Membranes, Beams and Plates, CRC Press, Taylor and Francis Group, Boca Raton, 2014.

[6] J. C. Hsieh and R. H. "Plaut, Free vibrations of inflatable dams," Acta Mech. 1990; vol. 85: pp. 207-220.

[7] M. C. Ece, M. Aydogdu and V. Taskin, "Vibration of a variable cross-section beam" Mech Res Commun. 2007, vol. 34, pp. 78-84.

[8] M. Şimşek, and T. Kocatürk, "Free vibration analysis of beams by using a third order shear deformation theory," Sadhana Acad Proc Eng Sci. 2007, vol. 32, pp. 167-179.

[9] Ö. Civalek, "Free vibration analysis of symmetrically laminated composite plates with first-order shear deformation theory (FSDT), by discrete singular convolution method," Finite Elem Anal Des. 2008, vol. 44, pp. 725-731.

[10] Ö. Civalek and M. Gürses, "Free vibration analysis of rotating cylindrical shells using discrete singular convolution technique," Int J Pres Ves Pip. 2009, vol. 86, pp. 677-683.

[11] M. Avcar, "Free vibration of randomly and continuously non-homogenous beams with clamped edges resting on elastic foundation," J Eng Sci Des. 2010, vol.1, pp. 33-38 (in Turkish).

[12] B. Akgöz and Ö. Civalek, "Longitudinal vibration analysis of strain gradient bars made of functionally graded materials," Compos B Eng. 2013, vol. 55, pp. 263-268. ·

[13] M. Avcar, "Free vibration analysis of beams considering different geometric characteristics and boundary condition," Int. Appl. Mech. 2014, vol.4, pp. 94-100.

[14] A. Prokic, M. Besevic and D. Lukic, "A numerical method for free vibration analysis of beams," Lat Am J Solid Struct. 2014, vol. 11, pp. 1432-1444.

[15] Y. Yeşilce, "Differential transform method and numerical assembly technique for free vibration analysis of the axial-loaded Timoshenko multiple-step beam carrying a number of intermediate lumped masses and rotary inertias," Struct Eng Mech. 2015, vol. 53, pp. 537-573

[16] L. A. Zadeh, "Fuzzy logic, neural networks and soft computing," Commun ACM. 1994, vol. 37, pp. 77-84.

[17] S. K. Das, A. Kumar, B. Das and A. P. Burnwal, "On soft computing techniques in various areas," Int J Inform Tech Comput Sci, 2013, vol. 3, pp, 59-68.

[18] H. Adeli, "Neural networks in civil engineering: 1989−2000," Comput Aided Civ Infrastruct Eng. 2001, vol. 16, pp. 126–142.

[19] R. Gates, M. Choi, S. K. Biswas and J.J. Helferty, "Stabilization of flexible structures using artificial neural networks," Proc Int Joint Conf Neural Network. 1993, vol. 2, pp. 1817-1820.

[20] B. Karlik, E. Özkaya, S. Aydin and M Pakdemirli, "Vibrations of a beam-mass systems using artificial neural networks," Comput. Struct. 1998, vol. 69, pp. 339-347.

[21] M. A. Mahmoud and M. A. A. Kiefa, "Neural network solution of the inverse vibration problem," NDT E Int. 1999, vol. 32, pp. 91-99.

[22] E. Castillo, A. Cobo, J. M. Gutierrez and E. Pruneda, "Functional networks: A new network–based methodology," Comput Aided Civ Infrastruct Eng.2000, vol. 15, pp. 90–106.

[23] M. Cevik, E. Özkaya and M. Pakdemirli, "Natural frequencies of suspension bridges: an artificial neural network approach," J Sound Vib. 2002, vol. 257, pp. 596-604.

[24] Ö. Civalek, “Flexural and axial vibration analysis of beams with different support conditions using artificial neural networks,” Struct Eng Mech. 2004, vol. 18, pp. 303-314.

[25] A. P. Hassanpour, E. Esmailzadeh and H. Mehdigholi, “Vibration of beams with unconventional boundary conditions using artificial neural network,” Proc ASME Int Des Eng Tech Conf Comput Inf Eng. 2005, vol. 1, pp. 159-165.

[26] S. M. Bagdatli, E. Özkaya, H. A. Özyiğit and A. Tekin, “Nonlinear vibrations of stepped beam systems using artificial neural networks,” Struct Eng Mech. 2009, vol. 33, pp. 15-30.

[27] R. A.. Saeed, A. N. Galybin and V. Popov. "Crack identification in curvilinear beams by using ANN and ANFIS based on natural frequencies and frequency response functions," Neural Comput. Appl. 2012, vol. 21 pp. 1629-1645.

[28] N. A. Jalil and I. Z. M. Darus, “Non-parametric Neuro-Model of a Flexible Beam Structure” IEEE Symp Comp Inf (ISCI), Langkawi, Malaysia, 2013, pp. 45-50.

[29] M. Mohammadhassani, H. Nezamabadi-pour, M. Suhatril and M. Shariati, “Identification of a suitable ANN architecture in predicting strain in tie section of concrete deep beams” Struct Eng Mech. 2013, vol. 46, pp. 853-868.

[30] Z. C. Ding, M. S. Cao, H. L. Jia, L. X. Pan and H. Xu “Structural dynamics-guided hierarchical neural-networks scheme for locating and quantifying damage in beam-type structures." J Vibroeng. 2014, vol. 16, pp. 3595-3608.

[31] M. Karimi, A. Shooshtaria and S. Razavia, “Large amplitude vibration prediction of rectangular plates by an optimal artificial neural network (ANN),” J Comput Appl Res Mech Eng. 2014, vol.4, pp. 55-65.

[32] Z. Şen, Principles of Artificial Neural Networks, Turkish Water Foundation Publications, 2004 (in Turkish).

[33] Ç. Elmas, Artificial Neural Networks (Theory, Architecture, Education, Application), Ankara: Seçkin Publication House, 2003 (in Turkish).

[34] S. Rajasekaran and G. Pai, Neural Networks, Fuzzy Logic And Genetic Algorithms, Synthesis & Applications. New Delh: Prentice-Hall of India Private Limited, 2003.

[35] M. M. Alshihri, A. M. Azmy and M. S. El-Bisy, “Neural networks for predicting compressive strength of structural light weight concrete,” Construct. Build. Mater. 2009, vol. 23, pp. 2214–2219.

[36] K. Saplıoğlu, M. Kilit and B. K. Yavuz, “Trend analysis of streams in the western Mediterranean basin of Turkey,” Fresen Environ Bull. 2014, vol. 23, pp. 313-324.

Fuzzy C-Means based Inference Mechanism for Association Rule Mining: A Clinical Data Mining Approach

Kapil Chaturvedi
Dept. of Computer Application, RGPV,
Bhopal, MP, India

Dr. Ravindra Patel
Dept. of Computer Application, RGPV
Bhopal, MP, India

Dr. D.K. Swami
Faculty of Engineering,
VNS Group of Inst.
Bhopal, MP, India

***Abstract*—Association rule mining has wide variety of research in the field of data mining, many of association rule mining approaches are well investigated in literature, but the major issue with ARM is, huge number of frequent patterns cannot produce direct knowledge or factual knowledge, hence to find factual knowledge and to discover inference, we propose a novel approach AFIRM in this paper followed by two step procedure, first is to discover frequent pattern by Appling ARM algorithm and second is to discover inference by adopting the concept of Fuzzy c-means clustering, for performance analysis, we apply this approach on a clinical dataset (contained symptoms information of patients) and we got highly effected disease in a couple of months or in a session as hidden knowledge or inference.**

***Keywords*—Association Rule Mining; Fuzzy Inference System; Clinical Data Mining; Preprocessing; Fuzzy clusters**

I. Introduction

Association rule mining (ARM) is the well-researched data mining technique [10, 15]. To find the frequent relation or association between items from market basket analysis perspective, which uses a rule based knowledge representation to find the relationship or causal dependencies between objects, things, attributes, outcomes, symptoms, occurrences, etc. It was first introduced in 1993 as the AIS algorithm [2] then in 1994 R. Agrawal and R. Srikant provided a candidate generation based technique formally known as Apriori algorithm [1] to generate rules, it is a streamlined version of AIS algorithm, it outperforms when support count is high and a number of items are less. The second approach for ARM is Frequent Pattern growth formally known as FP- Growth algorithm [11] proposed by J. Han, J. Pei and Y. in 2000.

But in present era we realize that association rule mining is a strong approach not only for market basket analysis rather than it plays vital role in various data analysis fields like stock market analysis, DNA pattern recognition, web data mining and also in clinical data mining, but the limitation of Association rule mining is it produces huge numbers of frequent patterns as per predefined thresholds which is insufficient to draw a conclusion.

The aim of this research work is to design and develop an inference mechanism for association rule mining, in order to discover abstract knowledge from the huge number of frequent patterns. Under this research, we proposed new algorithms, AFIRM to achieve the required goal. The major objectives of this research can be summarized in following points.

- To investigate the literature and problem identification.
- To choose a suitable approach for association rule mining for finding frequent patterns.
- To design an inference mechanism framework for association rule mining.
- To develop an inference mechanism for association rule mining in order to discover abstract/inference knowledge from frequent patterns.
- To analyze the result to meet abstract inference knowledge.

In this paper, we proposed a Fuzzy C-means (FCM) based inference system in order to discover inference knowledge, and the rest of the paper is organized as follows: section 2 elaborates the detail about fuzzy inference mechanism, section 3 give a comparison between hard and soft clustering approaches, section 4 introduces a literature review with basic terminology and previous research with related work already done in respective area, section 5 presented a comparative study between Apriori and FP-growth algorithm in order to choose most suitable ARM approach, section 6 proposed new approach along with illustrative example, simulation and result is presented in section 7, finally concluding remark is given in section 8.

II. Inference Mechanism

An inference engine is developed for an expert system consisting of inference mechanism as well as a control strategy. The term inference refers to the process of searching through the knowledge base and deriving new knowledge [18].

It involves formal reasoning by matching and unification, similar to the one performed by human experts to solve problems in a specific area of knowledge. An inference rule may be defined as a statement that has two parts an if clause and a then clause. It can be understood more clearly by adopting following example.

Rule 1: If Symptoms are headache, sneezing, running_nose then the patient have cold

Rule 2: if Symptoms are fever, cough and running nose, then patients have measles

III. Soft V/S Hard Clustering

Fuzzy c-means clustering algorithm is a little bit different from traditional clustering approaches due to its fuzzy nature and also because of the capacity of handling delicate data. It first introduced by Dunn [8] and then improved by Bezdek [3]. Traditional clustering algorithms (K-means [19], K-medoids [17]) are also known as hard clustering techniques because it divides the data in distinct clusters where each data element belongs to the exactly one group as shown in Figure-1 and 2.

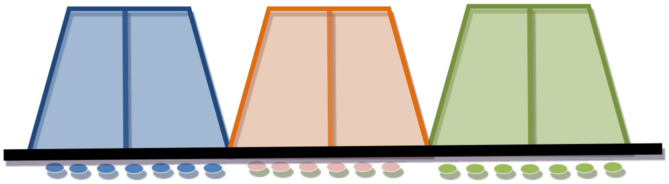


Fig. 1. Partitioning of data on the basis of similarity in traditional or Hard clustering approaches

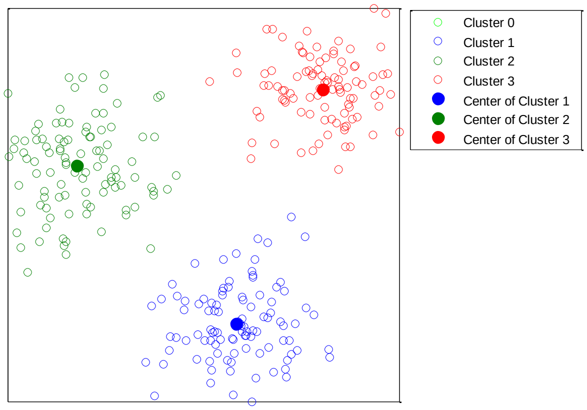


Fig. 2. Data clustering in traditional or Hard clustering approaches

Whereas fuzzy, c-means clustering or soft clustering algorithm provides an extended capacity of data categorization [7] where data elements can fit into more than one cluster and is associated with each object as shown Figure-3 & 4. This is a set of membership levels that specify the strength of the association between the data objects and a particular cluster. Fuzzy clustering is a process of assigning such membership levels and then using them to assign data objects to one or more clusters.

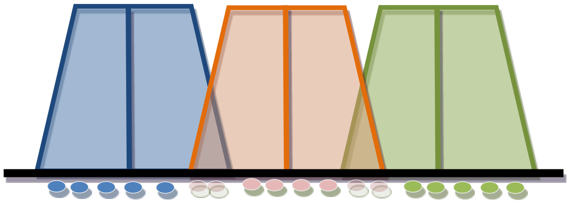


Fig. 3. Partitioning of data on the basis of degree of membership in fuzzy clustering approach

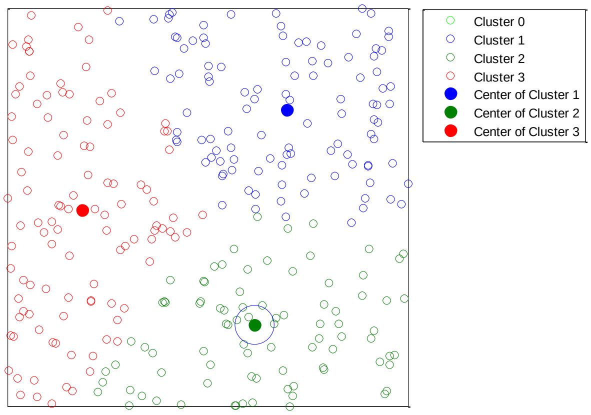


Fig. 4. Data clustering by fuzzy, c-means clustering approach

It is a strong approach from the data analysis perspective because it provides the way of organizing data into multiple predefined clusters or groups, based on their similarity among the classes from which it belongs to. But the additional extensibility of this approach is a common itemset that may refer to multiple clusters based on their degree of membership or similarity. Similarity refers to the mathematical calculation of similarity in the general distance calculation between objects using some well-defined distance measures.

A. Algorithm: Fuzzy C-Means (FCM)

Fuzzy c-means clustering algorithm is most suitable approach for implementing a fuzzy inference system, in this regard; we adopt the FCM algorithm to find factual knowledge hide in frequent patterns generated by ARM algorithm.

Input: D dimensional matrix, Number of Clusters – C

Output: Cluster Center Cq, Membership matrix U=[Uij]

Step-1: Let D dimensional matrix U with p data points represented by U=Upq, where initialize U=U (0)

Step-2: Assign the pre-defined number of clusters C, where 2≤C≤n

Step-3: At k-step calculate the center vector Cq, Where the cluster center C_q can calculate as follows:

$$C_q = \frac{\sum_{p=1}^{N} U_{pq}^{m} - X_p}{\sum_{r=1}^{C} U_{pq}^{m}}$$

Step-4: Calculate the degree of membership for pth data point in cluster q can be calculated as follows.

$$U_{pq} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| X_p - C_q \right\|}{\left\| X_p - C_r \right\|} \right)^{\frac{2}{m-1}}}$$

Step-5: Update Ur, Ur+1 (as STEP-4) at each step

Step-6: If $\left|U_{pq}^{r+1} - U_{pq}^{r}\right| < \varepsilon$, Then STOP

(Where, ε is the termination decisive factor or pre specified termination criteria between 0 and 1)

Step-7: Else return to Step-3

IV. Related Work

An inference mechanism framework for association rule mining proposed in [4] presents a theoretical and numerical study on association rule based inference mechanism for discovering factual knowledge from a clinical dataset in order to discover highly effected disease in a particular session by proposing an algorithm AIRM while this paper is presented a modified AIRM algorithm as AFIRM algorithm further describe in the next section.

In this manner to review the literature and to study on related work we investigate some good research approaches in the respective field, we also review the approaches for clinical data mining and fuzzy inference based approach.

A. *Inference approaches for ARM*

In [9] Ronald Fagin et al. presents a brief overview of inference rules, they also gave a brief discussion on the applicability of inference in various areas like inference in propositional logic, non-standard propositional logic, propositional modal logic and inference in first order logic, etc, X. Wanga et. al[23] proposed a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data, for this they made use of the cluster information generated by a self-organizing map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term (hourly) and long-term (daily) Web traffic trend predictions. Yang et al. [24] proposed an approach generic rule-base inference methodology using the evidential reasoning (RIMER) in this they proposed a new knowledge representation scheme in a rule base by analyzing existing knowledge base structure using a belief structure. Similarly R. Chow et. al [6] provides an ARM based Inference Mechanism, in this paper, they propose a refined and practical model of inference detection using a reference corpus. This model is inspired by association rule mining: inferences are based on word co-occurrences. Using the model and taking the Web as the reference corpus, that model also includes the important case of private corpora, to model inference detection in enterprise settings in which there is a large private document repository. They found inferences in private corpora by using analogues of his Web-mining algorithms, relying on an index for the corpus rather than a Web search engine. Tree-based mining approach for discovering patterns of human interaction in meetings [25] presented by Zhiwen Yu et. al for mining, human interactions for accessing and understanding meeting content for this they proposed a tree-based mining method for discovering frequent patterns of human interaction in meeting discussions. As per this study the mining results would be useful for summarization, indexing, and comparison of meeting records and interpretation of human interaction in meetings.

B. *ARM approaches for clinical data mining*

Data mining is more popular in the field of medical science due to its high applicability and analysis ability in medical and clinical data mining here we review some previous approaches related to medical field.

S. Venus et al. [21] proposed a rule based backward chaining inference engine which is an Arabic expert system based approach on natural language for diagnosing diseases, similarly Yanqing Ji et. al presents a potential causal association mining algorithm [16] for screening, adverse drug reactions in post marketing surveillance and proposed a novel data mining approach to signaling potential ADRs (Adverse drug reaction) from electronic health databases they also introduced potential causal association rules (PCARs) to represent the potential causal relationship between a drug and ICD-9 [13] coded signs or symptoms representing potential ADRs. Due to the infrequent nature of ADRs, the existing frequency-based data mining methods cannot effectively discover PCARs. They introduce a new interesting measure, potential causal leverage, to quantify the degree of association of a potential causal association rule (PCAR) similarly paper [20, 22] proposed practical and applied fuzzy logic based approaches for medical diagnosis, whereas paper [5] presents a novel data mining approach to generate adverse drug events detection rules. The main objective of this paper is to automatically detect cases of ADEs (adverse drug events) by data mining. They used decision trees and association rules to discover ADE detection rules, with respect to time constraints. The rules are then filtered, validated, and reorganized by a committee of experts. The rules are described in a rule repository, and several statistics are automatically computed in every medical department, such as the confidence, relative risk, and median delay of outcome appearance.

V. Performance Analysis of Apriori & Fp-Growth Algorithms

In this section, performance analysis has been carried out in order to analyze the efficiency of Apriori and FP-Growth algorithm moreover choose the most appropriate ARM algorithm for finding frequent patterns, for this purpose we apply the both algorithms on T10I4D100K [12] and pima D38.N768.C2 [14] datasets, and measure time efficiency on different support counts, as shown in figure 5 and 6. Above analysis can be outlined in following points.

- FP-Growth algorithm outperforms on low support count.
- Apriori performs better on the high support count.
- FP-Growth takes much processing time, if large transaction set is given, due to its requirement of large storage memory in order to store a tree structure.

TABLE I. EXECUTION TIME OF APRIORI AND FP-GROWTH ON DIFFERENT SUPPORT COUNTS (FOR T10I4D100K DATASET)

Support Count	**Time (MS)**	
	FP-Growth	Apriori
20	26900	46907
30	3737	7860
40	1206	2301
50	1126	929
60	1009	604
70	973	613

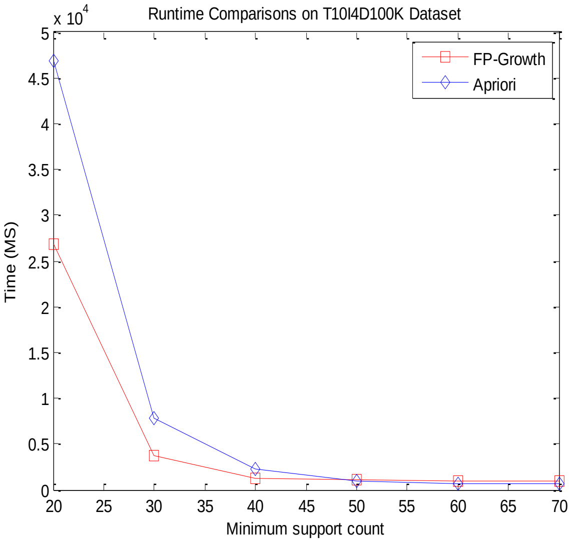


Fig. 5. Runtime comparative analyses on T10I4D100K Dataset

TABLE II. EXECUTION TIME OF APRIORI AND FP-GROWTH FOR PIMA D38.N768.C2 DATASET

Support Count	**Time(MS)**	
	FP-Growth	Apriori
10	51	94
20	42	78
30	42	72
40	35	55
50	30	45
60	25	32
70	19	25
80	16	16
90	8	7

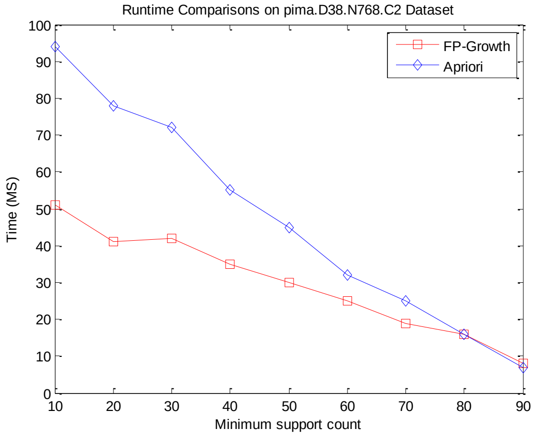


Fig. 6. Runtime comparative analyses on pima.D38.N768.C2 Dataset

VI. PROPOSED APPROACH

To overcome the limitations of the previous AIRM approach we adopt the concept of fuzzy clustering instead of classification used in previous (AIRM) approach which accepts frequent patters with degree of matching 1 (100%), whereas the newly proposed algorithm accepts all the frequent patterns either it has degree of matching 0, 1 or between of 0 & 1 because of its fuzzy nature. Under this process FCM accepts the normalized data set which contained normalized degrees of matching and produces clusters to demonstrate the nature of frequent patterns. In order to implement a fuzzy inference mechanism as shown in figure.7, AFIRM follows the following steps.

A. Data selection and Preprocessing

This is the very first step of AFIRM algorithm where data selection has performed for preprocessing, this step first load and scan the main dataset as well as fact dataset, then preprocess the datasets by mapping items/ symptoms by its corresponding index values. For this it previously read all items/symptoms and create an index table by assigning a unique index value to each item/symptom.

B. Association rule mining

This step performs the association rule mining in order to find frequent patterns, hence we use the FP - growth algorithm because it reads database once, and find frequent patterns. This process generates all the frequent patterns or frequent symptoms as per given thresholds.

C. Fact matching and matrix generation

This is the most significant step of this algorithm where all the frequent symptoms match with fact dataset and find its degree of similarity (where factual data set contain symptoms with related disease), afterward all these information further stores in another database where frequent symptoms, disease and degree of similarity camps to gather in the form of a matrix.

D. Matrix normalization

Once the matrix of DOM is completely formed, a normalization step is conceived to normalize the degree of similarity for applying fuzzy, c-means algorithm, here all the degrees must be between 0 and 1. If the degree is 0 means no similarity is found and if degree is 1 means frequent symptom fully similar to disease symptoms and if the degree is between 0.01 to 0.99 mean partial similarity has found.

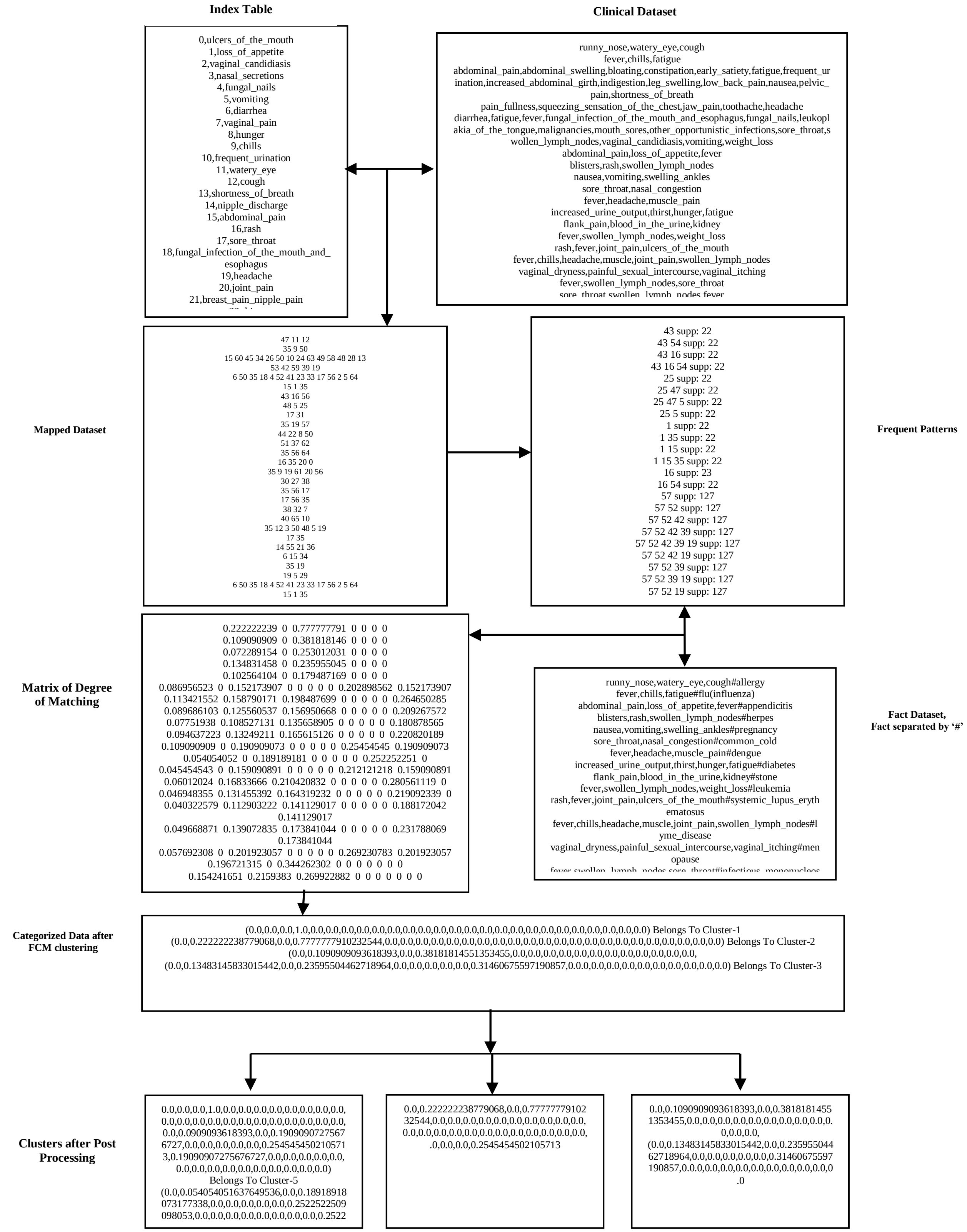


Fig. 7. Proposed Framework for AFIRM

E. FCM Clustering and Inference evaluation

FCM is the fuzzy based clustering algorithm so that it can better categorize the above generated degrees of similarity, hence we apply the Fuzzy c-means algorithm on normalized matrix in order to categorize the frequent symptoms. When we plot these clusters, we seem that visualization, demonstrate that which diseases is highly effected in a session.

VII. ALGORITHM: ASSOCIATION FUZZY INFERENCE RULE MINER (ABBREVIATED AS AFIRM)

TABLE III. ABBREVIATIONS USED:

Abbreviation	Meaning
DS	Dataset
FP	Frequent patterns
ARM	Association rule mining
Infr	Inference
Min_sup	Minimum support
FIS	Fuzzy inference system
FCM	Fuzzy C-Means
DMMTX	Degree of Matching Matrix
NDMMTX	Normalize Degree of Matching Matrix

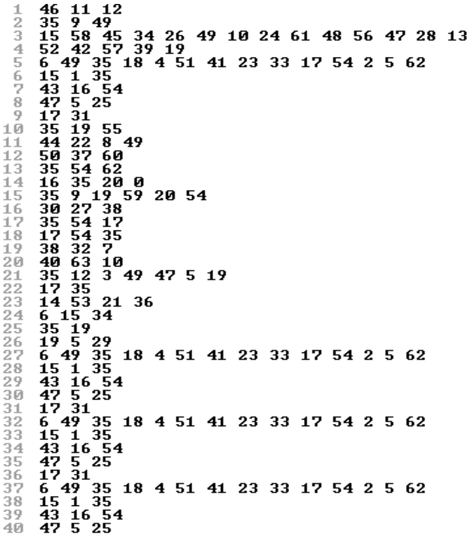
```
1.  Begin:
2.  Load (Dataset), Min_Sup_Threshold, Fact_DS
    //Data Preprocessing
3.  Index=Gen_Index(Dataset)
4.  Mapped_DS=Mapping(Dataset, Index)
    //Association Rule Mining
5.  FP=ARM(Mapped_DS, Min_Sup_Threshold)
    // FP-Growth By Han et.al [20]
6.  DMMTX=Match(FP, Fact_DS)
    // Fact Matching & DOM Matrix Implementation
7.  NDMMTX=Normalization(DMMTX)
    //Normalization of DOM matrix
8.  Infr_Clusters=FCM(FP, Fact_DS)
9.  PostProcessing(Infr_Clusters)
10. Result_Analysis
11. End
```

Procedure: FCM (NDMMTX)

```
1. Begin
2. Initialize matrix U for degree of membership U=[uij]
3. At each step calculate centers vector C (k) & U (k).
4. Calculate the d-dimension center of the cluster.
5. Update U(k), U(k+1)
6. Repeat step 2 till MAXij< termination decisive factor
7. End.
```

A. Case Study

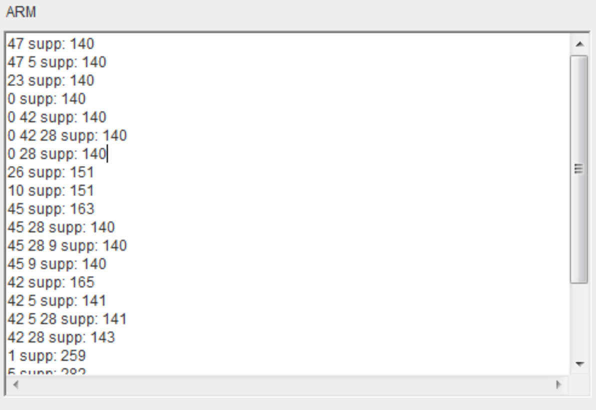
After preprocessing, original dataset will map in following form.



```
46 11 12
35 9 49
15 58 45 34 26 49 10 24 61 48 56 47 28 13
52 42 57 39 19
6 49 35 18 4 51 41 23 33 17 54 2 5 62
15 1 35
43 16 54
47 5 25
17 31
35 19 55
44 22 8 49
50 37 60
35 54 62
16 35 20 0
35 9 19 59 20 54
30 27 38
35 54 17
17 54 35
38 32 7
40 63 10
35 12 3 49 47 5 19
17 35
14 53 21 36
6 15 34
35 19
19 5 29
6 49 35 18 4 51 41 23 33 17 54 2 5 62
15 1 35
43 16 54
47 5 25
17 31
6 49 35 18 4 51 41 23 33 17 54 2 5 62
15 1 35
43 16 54
47 5 25
17 31
6 49 35 18 4 51 41 23 33 17 54 2 5 62
15 1 35
43 16 54
47 5 25
```

Fig. 8. Mapped dataset

After applying ARM we got 511 frequent symptoms on 40% of minimum support.



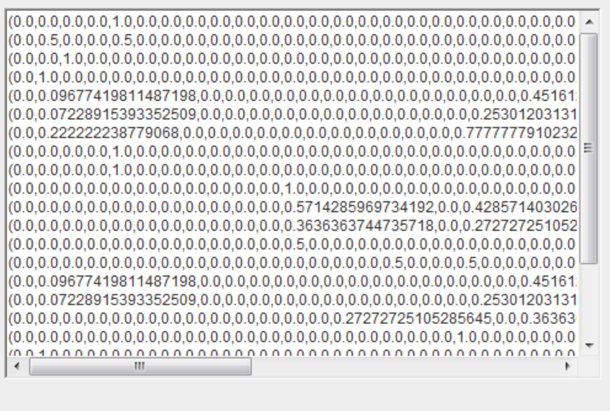
ARM

```
47 supp: 140
47 5 supp: 140
23 supp: 140
0 supp: 140
0 42 supp: 140
0 42 28 supp: 140
0 28 supp: 140
26 supp: 151
10 supp: 151
45 supp: 163
45 28 supp: 140
45 28 9 supp: 140
45 9 supp: 140
42 supp: 165
42 5 supp: 141
42 5 28 supp: 141
42 28 supp: 143
1 supp: 259
5 supp: 282
```

Fig. 9. Frequent pattern discovery under ARM

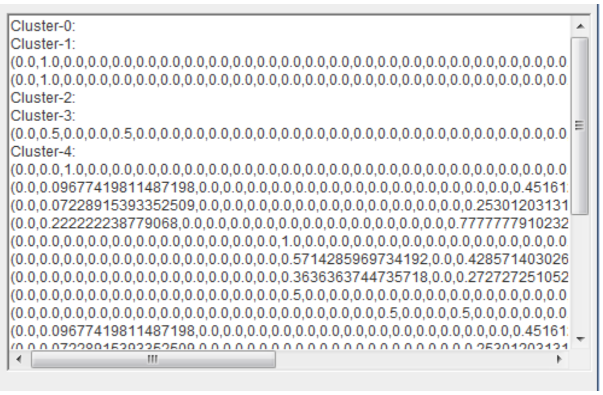
Figure.11 shows the resulting clusters, which categorize the frequent symptoms on the basis of degree of matching (DOM) in order to demonstrate and categorize the possibility of disease infection. Earlier than, DOM normalization has applied to transform the DOM in the required form, as shown in Figure-10.

After clustering a post processing phase is needed to cluster the data and validate the result, it depends on data properties, where data points are plotted with different colors depending on the cluster assigned. In this step we apply the post processing on clusters, to obtain the exact information about clusters.



```
(0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.5,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.09677419811487198,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.45161
(0.0,0.07228915393352509,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25301203131
(0.0,0.222222238779068,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7777777910232
(0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5714285969734192,0.0,0.428571403026
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.3636363744735718,0.0,0.272727251052
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.5,0.0,0.0,0.0,0.0
(0.0,0.09677419811487198,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.45161
(0.0,0.07228915393352509,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25301203131
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.27272725105285645,0.0,0.36363
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0
(0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
```

Fig. 10. Normalized (Pre-processed) matrix for FCM



```
Cluster-0:
Cluster-1:
(0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
Cluster-2:
Cluster-3:
(0.0,0.5,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
Cluster-4:
(0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.09677419811487198,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.45161
(0.0,0.07228915393352509,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25301203131
(0.0,0.222222238779068,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.7777777910232
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5714285969734192,0.0,0.428571403026
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.3636363744735718,0.0,0.272727251052
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
(0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.5,0.0,0.0,0.5,0.0,0.0,0.0,0.0
(0.0,0.09677419811487198,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.45161
(0.0,0.07228915393352509,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.25301203131
```

Fig. 11. Clusters generated by FCM algorithm

VIII. Simulation and Result

Figure 12 represents a graphical representation of clusters generated by fuzzy, c-means algorithm, this scatters view, representing the amount of possible disease in a particular cluster with the respective degree of matching. In this figure cluster 2 and 4 contain the high number of objects as diseases and showing that the disease5 has a high possibility of infection under this study, we also infer that the cluster 2, 4 and 5 are represented which disease would be highly effective.

IX. Conclusion

In this paper, we have proposed a fuzzy, c-means based inference system for association rule mining, in this study, we propose a novel algorithm AFIRM (Association fuzz inference rule mining) which is an extended version of previously proposed algorithm AIRM.

It consists of two phases. First phase scan the given dataset with corresponding fact dataset and perform preprocess to meet required format for rule mining then we apply FP-growth on pre-processed dataset in order to find frequent patterns. Second, we match these frequent patterns with fact dataset and create a matrix of degree of matching or similar and then, we normalize this matrix to apply FCM procedure. Therefore Fuzzy c-means clustering algorithm categorizes this data in different clusters on the basis of pre assigned degree of matching. Experimental results show that FIRM (Fuzzy inference rule miner) out performs comparisons to previous approach AIRM in order to discover background knowledge more over use full inference.

In future AFIRM can also use the concept of Markov predictor to know future possibilities. Secondly AIRM and AFIRM both algorithms first discover the association rules using FP-growth algorithm, to optimize, an expert mechanism might be explored to find most suitable and efficient algorithm instead of FP-growth algorithm on the basis of the type of dataset to efficiently discover association rules.

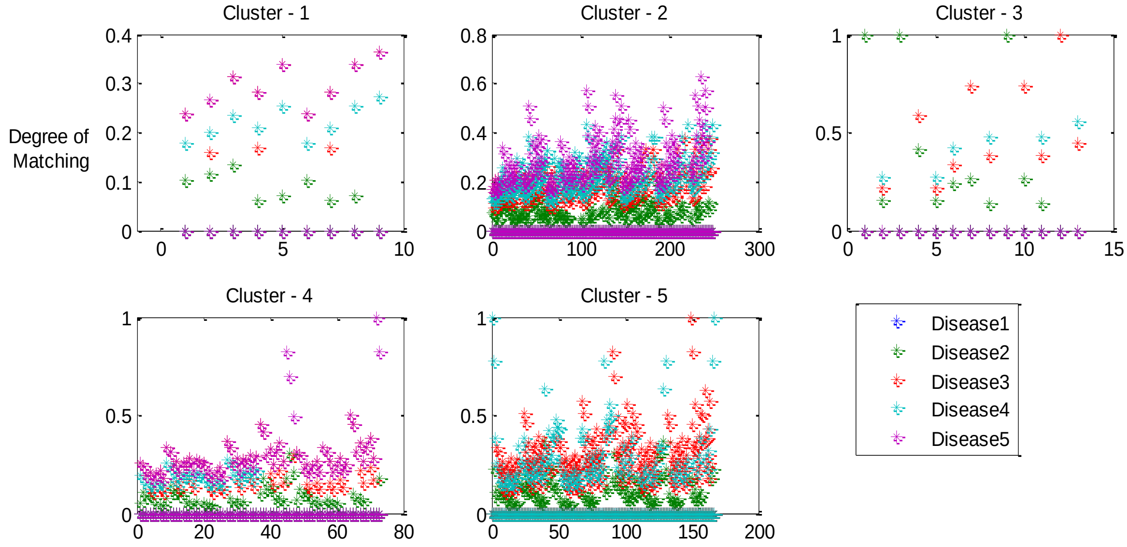


Fig. 12. Clusters generated after FCM, representing amounts of possible diseases in resulting clusters

REFERENCES

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." In Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, (1994): pp. 487-499.

[2] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." In ACM SIGMOD Record, vol. 22(2), (1993): pp.207-216.

[3] Bezdek, James C., "Pattern recognition with fuzzy objective function algorithms", Kluwer Academic Publishers, 1981.

[4] Chaturvedi, Kapil, Dr. Ravindra Patel and Dr. D.K. Swami, "An Inference Mechanism Framework for Association Rule Mining" International Journal of Advanced Research in Artificial Intelligence(IJARAI), vol. 3(9), (2014): pp.1-8.

[5] Chazard, Emmanuel, Gregoire Ficheur, Stephanie Bernonville, Michel Luyckx, and Regis Beuscart. "Data mining to generate adverse drug events detection rules." Information Technology in Biomedicine, IEEE Transactions, vol. 15(6), (2011): pp.823-830.

[6] Chow, Richard, Philippe Golle, and Jessica Staddon. "Detecting privacy leaks using corpus-based association rules." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, (2008): pp. 893-901.

[7] Cominetti, Ornella, Anastasios Matzavinos, Sandhya Samarasinghe, Don Kulasiri, Sijia Liu, Philip Maini, and Radek Erban. "DifFUZZY: a fuzzy clustering algorithm for complex datasets." International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol. 1(4), (2010): pp. 402-417.

[8] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." Journal of cybernetics, vol. 3, (1973): 32-57.

[9] Fagin, Ronald, Joseph Y. Halpern, and Moshe Y. Vardi. "What is an inference rule?." The Journal of symbolic logic, vol. 57(03), (1992): pp.1018-1045.

[10] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine, vol. 17(3), (1996): pp.1-34.

[11] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." ACM SIGMOD Record, vol. 29(2), (2000).

[12] http://fimi.ua.ac.be/data/T10I4D100K.dat (Frequent Itemset Mining Dataset Repository).

[13] http://www.cdc.gov/nchs/icd/icd9.html

[14] http://www.cs831/notes/itemsets/datasets.php (the lucs-kdd software library)

[15] J. Han, M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, USA, ISBN 1558604898, 2001.

[16] Ji, Yanqing, Hao Ying, Peter Dews, Ayman Mansour, John Tran, Richard E. Miller, and R. Michael Massanari. "A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance."Information Technology in Biomedicine, IEEE Transactions, vol. 15(3), (2011): pp.428-437.

[17] Kaufman, Leonard, and Peter J. Rousseeuw. "Partitioning around medoids (program pam)." Finding groups in data: an introduction to cluster analysis (1990): pp. 68-125.

[18] Kaushik, Saroj, "Artificial Intelligence", Cengage Learning, India, ISBN.9788131510995, 2011.

[19] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, (1967): pp. 281-297.

[20] Nguyen, Thanh, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi, "Medical diagnosis by fuzzy standard additive model with wavelets." Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on. IEEE, (2014): pp.1937-1944.

[21] Samawi, Venus , Akram Mustafa, and Abeer Ahmad. , "Arabic Expert System Shell", IAJIT First Online Publication, (2011): vol. 10(1).

[22] Tao, Xuehong, Yuan Miao, Yanchun Zhang, and Zhiqi Shen. "Collaborative medical diagnosis through Fuzzy Petri Net based agent argumentation." InFuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on, (2014): pp.1197-1204.

[23] Wang, Xiaozhe, Ajith Abraham, and Kate A. Smith. "Intelligent web traffic mining and analysis." Journal of Network and Computer Applications, Vol. 28(2), (2005): pp.147-165.

[24] Yang, Jian-Bo, Jun Liu, Jin Wang, How-Sing Sii, and Hong-Wei Wang. "Belief rule-base inference methodology using the evidential reasoning approach-RIMER." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions, vol. 36(2), (2006): pp.266-285.

[25] Yu, Zhiwen, Zhiyong Yu, Xingshe Zhou, Christian Becker, and Yuichi Nakamura. "Tree-based mining for discovering patterns of human interaction in meetings." Knowledge and Data Engineering, IEEE Transactions, vol. 24(4), (2012): pp.759-768.

AUTHOR PROFILE

Mr. Kapil Chaturvedi has post graduated in Master of Computer Application (MCA) from from Samrat Ashok Technical Institute (Degree), Vidisha, India. Presently he is pursuing Ph.D. from Rajiv Gandhi Technological University, (Bhopal) India. He has more than 6 year teaching experience in post-graduation, He has published more than 6 research papers in international and conference proceedings He is working as Assistant Professor in Department of Computer Applications, UIT, Rajiv Gandhi Technical University, Bhopal, India, his area of interest is Data Mining & Digital Image Processing.

Dr. Ravindra Patel, Associated Professor and Head, Department of Computer Applications at Rajiv Gandhi Technological University, (Bhopal) India. He is Ph.D. in Computer Science. He possesses more than 10 years of teaching experience in post-graduation. He has published more than 30 research papers in international and national journals, and conference proceedings. He is member of ISTE and International Association of Computer Science and Information Technology (IACSIT).

Dr. D. K. Swami is presently Group Director, VNS Group of Institutions, Bhopal. Before joining VNS in 2008, he has served Samrat Ashok Technological Institute (SATI), Vidisha for 20 years. He obtained his Ph. D. in Computer Science and Engineering in 2007 from Rajiv Gandhi Technological University, Bhopal. He completed M. Tech in Computer Applications from IIT, Delhi in 1994 and he completed M.Sc. in Applied Mathematics in 1988. During 26 years of teaching, he has taught various courses in computer science to the students of BE, M. Tech. and MCA students. He has more than 30 research publications I journals and in conference proceedings of repute. He has edited one book and authored a book on Basic Computer Engineering. He is a member for life ISTE and senior life member of CSI. Areas of his special interest include Data Mining, DBMS, Object Oriented System and software Engineering.

Integration of Qos Aspects in the Cloud Service Research and Selection System

[1]Manar ABOUREZQ and [2]Abdellah IDRISSI
Computer Sciences Laboratory (LRI), Computer Sciences Department
Faculty of Sciences, Mohammed V University of Rabat

Abstract—**Cloud Computing is a business model revolution more than a technological one. It capitalized on various technologies that have proved themselves and reshaped the use of computers by replacing their local use by a centralized one where shared resources are stored and managed by a third-party in a way transparent to end-users. With this new use came new needs and one of them is the need to search through Cloud services and select the ones that meet certain requirements. To address this need, we have developed, in a previous work, the Cloud Service Research and Selection System (CSRSS) which aims to allow Cloud users to search through Cloud services in the database and find the ones that match their requirements. It is based on the Skyline and ELECTRE IS. In this paper, we improve the system by introducing 7 new dimensions related to QoS constraints. Our work's main contribution is conceiving an Agent that uses both the Skyline and an outranking method, called ELECTREIsSkyline, to determine which Cloud services meet better the users' requirements while respecting QoS properties. We programmed and tested this method for a total of 10 dimensions and for 50 000 cloud services. The first results are very promising and show the effectiveness of our approach.**

Keywords—Cloud Computing; Cloud Services; Quality of Service; Skyline; Outranking methods; Multi criteria decision; ELECTRE methods; Block-Nested Loops Algorithm

I. INTRODUCTION

Cloud Computing refers to software, hardware and datacenters offered as a service over a network and remotely accessible via various devices such as computers, PDAs, smart phones, etc. Although it is a rather new computing paradigm that appeared in the last decade, Cloud Computing capitalizes on concepts that have been proven, such as Virtualization [1], Distributed Computing [2], Grid Computing [3], Web Services [4], Service-Oriented Architecture [5], etc.

One of the early definitions of Cloud Computing was proposed by Wang et al. [6], who defined Cloud Computing as *"a set of network enabled services, providing scalable, QoS guaranteed, normally personalized, inexpensive computing platforms on demand, which could be accessed in a simple and pervasive way"*.

Another definition based on the concepts Cloud Computing is built on was proposed by Vouk in [7], stating that Cloud Computing "*embraces cyber infrastructure and builds upon decades of research in virtualization, distributed computing, grid computing, utility computing, and, more recently, networking, web and software services. It implies a service-oriented architecture, reduced information technology overhead for the end-user, greater flexibility, reduced total cost of ownership, on-demand services and many other things*".

In [8], Cloud Computing is defined as being a *"type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers"*.

The NIST [9] defines Cloud Computing as being *"a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction"*.

Foster et al. propose another definition in [3] where Cloud Computing is considered as *"a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet"*.

Though there are many attempted definitions of Cloud Computing, they all agree that every Cloud system has the following essential characteristics:

- The use of virtualization to offer a set of shared physical and virtual resources such as networks, servers, storage space, bandwidth, applications…;
- Dynamic configurability that makes it easy to expand or decrease depending on the user's needs, without affecting the level of reliability and security;
- Accessibility via a network, usually the Internet, from various machines (computers, smart phones, tablets, PDAs…) using standard APIs;
- The use of specific measure systems to control and optimize the use of resources and to offer a billing based on what was consumed, without surplus or need of managing the underlying infrastructure.

The services reachable via Cloud may be divided into three categories [10]: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Each one of these categories has specific characteristics that make it more adapted to certain user groups. For instance, enterprises will more likely purchase IaaS and PaaS services, while individuals will be more inclined to use SaaS services.

SaaS [11] allows users to remotely access applications that run in the Cloud's infrastructure by using thin or thick clients. Thus, there is no need to invest in an infrastructure or to buy software licenses. For providers, costs of installation, hosting and maintenance are optimized since many users access to the same application. Examples of SaaS include Google Drive [12] (formerly Google Docs) and Salesforce CRM [13].

PaaS [14] offers a software layer or a development environment as a service on which users will build and deploy their own applications. That way, users won't need to manage the infrastructure while keeping control of the deployed applications and configuring the hosting environment. Examples of PaaS include Salesforce's Force.com [15], Google App Engine [16] and Microsoft Windows Azure [17].

IaaS [18] provides as a service basic storage and computing resources such as servers, network equipments, data warehouses… These resources will be used to run users' own applications. Usually, IaaS satisfies best the end-users' needs of interoperability and portability [19] because they choose the various blocks that compose the infrastructure used. Examples of IaaS include Amazon Elastic Compute Cloud [20] and Microsoft SQL Azure [21].

In addition to these three main models, many others have been proposed such as Hardware as a Service [6], Communication as a Service [22], Network as a Service [23], Data as a Service [22], Workplace as a Service [24], Security as a service [25], Business Process as a service [26], Identity and Policy Management as a Service [27], STorage as a Service [28], Cluster as a Service [29], etc.

Cloud services can be deployed in various models [30], depending on the use case, the provider's business model... The most widespread deployment models are Public, Private, Community and Hybrid.

A Public Cloud [30] is an open Cloud provided by an organization to the general public. It can be accessed via a network, usually the Internet. However, the fact that the Cloud is public doesn't imply that services are offered for free or that the data exchanged by its means is not confidential.

A Private Cloud [30] is offered to the sole use of one organization that either manages it or delegates its management to a third-party. The main advantage of this deployment model is that there are no limitations regarding bandwidth or security, since the resources are exclusively used by the organization.

A Community Cloud [19] is a Cloud shared by organizations belonging to the same community. They can manage their Cloud themselves or delegate the chore to a third-party.

A Hybrid Cloud [31] contains two or more of the Cloud models cited above interconnected by standard or proprietary technologies.

In addition to these four deployment models, new ones are emerging, like the On-Site Private Cloud [19] and the Special Purpose Cloud [32].

The On-Site Private Cloud is a Cloud intended for the private use of a sole organization, just like the Private Cloud. However, it is hosted by the organization, either in a centralized or distributed way. The security aspect is also managed by the organization.

The Special-Purpose Cloud provides, on top of standard resources, additional methods regarding specific use cases. An example that illustrates this model is Google's App Engine with the specific capacities it offers to document management.

Using a Cloud service presents many advantages to end-users. First, there is a significant cost reduction, since users purchase only the resources they need, without surplus or need to invest in infrastructure or maintenance. There's also the guarantee of instant and uninterrupted access to computing and storage resources to any user who has a network connected machine. In addition to it, users can easily adapt the available resources to their specific needs and can add resources as required.

All these advantages have led to an increase in the use of Cloud Computing. With this increase, many new needs have emerged, among which there is the need to find Cloud services that match the users' requirements. Our contribution is in this research area and consists of a Cloud Services Research and Selection System (CSRSS) based on the Skyline and ELECTRE IS as presented in [35] and [36].

The CSRSS allows users to specify the technical and functional requirements of the cloud services they want to use. To do so, it connects to a database of Cloud services and selects the ones that match the users' requirements while giving them the possibility of getting the optimal value of some of these requirements.

With the CSRSS returning the Cloud services that satisfy best the technical and functional requirements specified by users, our objective now is to address Quality of Service (QoS) requirements to better refine the resulting services by keeping the ones that best satisfy QoS parameters.

Our paper is organized as follows. In the next section we present some related work. In section 3, we present the Cloud Service Research and Selection System (CSRSS) as presented in [36]. In section 4, we present QoS aspects and how we integrated them into the CSRS System. In section 5, we present our improved prototype and test results before concluding in section 6.

II. Related Work

The increase use of Cloud Computing has resulted in the emergence of new needs, such as the need of having systems to search and select Cloud services that meet users' requirements. Many works have been carried out to offer new solutions that will help users to choose the Cloud services that answer their needs. It is rather different from the selection of Cloud service components for composition purpose, which is beyond the scope of our research subject. Our main goal is to find Cloud services that best match the users' requirements, not the selection of two or more Cloud services to compose one final Cloud service that will be delivered to users.

One of the first works dealing with Cloud services discovery and selection was proposed by Goscinski et al. in [29]. It focused on Cloud clusters and used a broker that dynamically matches services and clusters.

Zeng et al. presented a Cloud service selection algorithm in [37]. The algorithm determines the cost and gains of available Cloud services that can be reached via proxy and return those that maximize the gains and minimize the cost. It is done in two steps. In the first step, the proxy selects the available Cloud services following the request sent by the user. In the second step, the algorithm computes the gains and cost of the selected Cloud services and returns the ones that optimize the cost and gains.

Kang and Sim presented a Cloud portal with a Cloud service search engine in [38]. This system uses the concept of similarity [39] and consults the adopted Cloud ontology to select the Cloud services that match the requirements specified by the user.

Kang and Sim also proposed Cloudle in [40], a Cloud services search engine which main functionalities are query processing, similarity reasoning and rating. Like the portal presented in [38], Cloudle consults a Cloud ontology to compute the similarity between Cloud services and returns a list of results sorted by aggregated similarity.

In [41], Han and Sim presented a Cloud Service Discovery System. It consults a Cloud ontology to compute the similarity between Cloud services and return a list of results matching the user's query.

In the three systems presented above, users specify the requirements that must be satisfied by the Cloud services they are searching for. These requirements can be split into three types, namely functional requirements (category of service), technical requirements (OS, CPU, memory, storage space...) and cost requirements (price and timeslot range).

Yoo et al. presented in [42] a resource selection service based on Cloud ontology. It generates Virtual Ontologies (Vons) based on virtualized resources and combine them into new resources. Then it computes the similarity between these new resources to determine the ones that meet best the user's requirements.

In [43], Zang et al. presented a service matching algorithm and a service composition algorithm. These algorithms search through Cloud services and compute the semantic similarity [39] between them to determine whether two given Cloud services are interoperable.

These systems mostly use similarity [39] to determine which Cloud service is the most similar to the user's quest. Thus, they would be better suited for users who want to find Cloud services that are similar to the ones they already know or use. This leaves out users that want to find Cloud services that meet some requirements (service model, provider, bandwidth, latency...) without knowing an existing Cloud services that does meet these requirements. This is why we have thought of replacing similarity with the principle of the Skyline [33].

Using the Skyline allows users to specify the criteria they want to optimize and to get the Cloud services that meet their needs. Thereby, we have developed in [35] a system that enables them to do so and that is based on the principle of the Skyline. We then tried to improve our system by applying outranking methods [34] to the results returned by the Skyline.

There are many works that have used MCDM methods to address the selection of Cloud services. L. Sun et al. conducted a thorough study of Cloud service selection techniques in [44], including MCDM-based techniques such as Analytic Hierarchy Process (AHP) [45], Analytic Network Process (ANP) [46] MAUT [47] and outranking methods [34].

Garg et al. presented a framework for ranking Cloud services based on AHP in [48]. They used the metrics that form the Service Measurement Index (SMI) [49] to define the criteria upon which Cloud services are to be evaluated and compared. They chose six criteria which are accountability, agility, assurance, cost, performance, and security. Although this method is interesting, it has only been tested using three Cloud services as an input in one use case, and 1000 providers in the other.

Godse et al. proposed in [50] an AHP-based approach for the selection of SaaS products Cloud services. The criteria used are functionality, architecture, usability, vendor reputation, and cost. Like the previous one [48], this method has only been tested using three Cloud services (SaaS products). Also, it is only used to compare SaaS Cloud services, leaving out other Cloud services categories.

Karim et al. presented in [51] a QoS mapping approach for combining SaaS and IaaS products and then ranking the combined Cloud services for end users. This method is carried out in four mapping steps. First, the users' QoS requirements are mapped to the QoS specifications of available SaaS products. Second, the obtained SaaS products are mapped to the available IaaS products that have the best QoS guarantees. Third, the end-to-end resulting QoS specifications are computed. Finally, AHP is used to rank the combined Cloud services based on the end-to-end QoS specifications obtained. Tests have been carried using four Cloud services and eight QoS criteria. It is also intended to be used for IaaS and SaaS Cloud services only.

Menzel et al. present in [52] a Multi-Criteria Comparison Method for Cloud Computing, denoted $(MC^2)^2$,that offers a framework for selecting Cloud services using ANP, which is an extension of AHP proposed by Saaty in [46]. This framework allows users to select the best adapted IaaS to their needs. Nine criteria are used such as flexibility to change, reliability, security, maturity of the provider…The $(MC^2)^2$ has been implemented as a web application; AOTEAROA.

Limam and Boutaba proposed in [53] a Cloud service selection approach based on MAUT and aimed at SaaS products. In order to use the MAUT method, the three criteria initially chosen, namely reputation, quality, and cost are reduced to one criterion; feedback.

Silas et al. presented in [54] a middleware for the selection of Cloud services using ELECTRE. Like most of the works cited above, the criteria used are QoS related. The middleware uses ELECTRE III to rank the Cloud services according to the degree to which they match the user's requirements and preferences.

To our knowledge, no work has combined the use of the Skyline operator and ELECTRE. Our motivation to do so in [36] is 1) to capitalize on the results of our first work [35] and refine its results and 2) to minimize the complexity that comes from using ELECTRE alone. Indeed, ELECTRE carries a pairwise comparison to build a decision matrix which size is n x n, n being the number of alternatives. The prior use of the Skyline operator allows making a first filtering of the input, reducing its size up to more than 40%. Thus, the alternatives that are contained in the Skyline form the input to the ELECTRE algorithm, knowing that the Skyline contains all the interesting alternatives for the user, no matter how they weight their preferences. In other words, we apply ELECTRE to less than 60% of the candidate alternatives, after eliminating the rest using the Skyline operator.

We present the prototype and algorithms of our system, as presented in [36], in the next section.

III. THE CLOUD SERVICE RESEARCH AND SELECTION SYSTEM (CSRSS)

The figure below illustrates the prototype of our CSRS System as presented in [36]. It involves the introduction of several agents and consists of a user interface, a user's query processing agent, a pre-Skyline processing agent, a cloud services research and selection agent, an ELECTRE IS agent and a database.

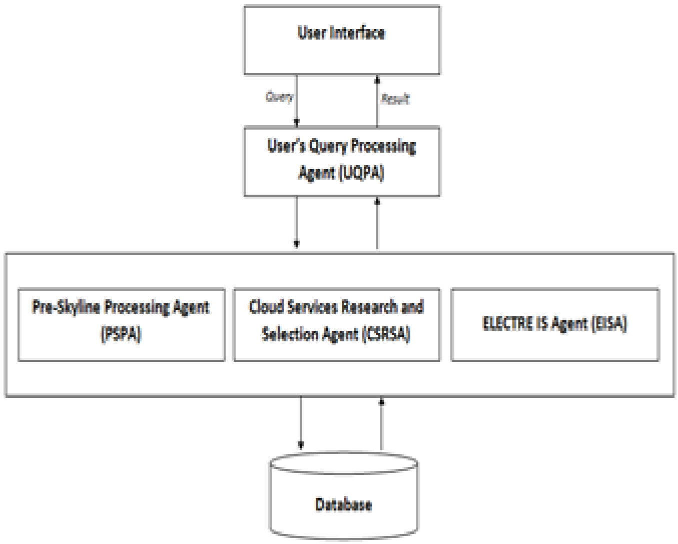

User Interface
Query
Result
User's Query Processing Agent (UQPA)
Pre-Skyline Processing Agent (PSPA)
Cloud Services Research and Selection Agent (CSRSA)
ELECTRE IS Agent (EISA)
Database

Fig. 1. A schema representing the new version of the Cloud Service Research and Selection System (CSRSS)

The user's interface allows users to interact with the system by selecting the requirements that the Cloud services must meet and view the returned results. It also allows the users to add Cloud services by filling in their attributes such as the name, the provider, the bandwidth, the OS, etc. We think that these requirements are the common ground to existing and upcoming Cloud ontologies [22, 38, 40, 41, 42, 55].

The user's query processing agent extracts the requirements contained in the user's request and sets them into two categories:

- Requirements that are fixed, such as the provider's name, the service model, the OS…;
- Requirements that are to be optimized, such as the price (to be minimized), the bandwidth (to be maximized), etc. These requirements will be used as the Skyline's dimensions.

The Cloud Services Research and Selection Agent (CSRSA) connects to the database and executes a SQL query, which predicates are the fixed requirements returned as a result by the user's query processing, to select all the Cloud services that meet these fixed requirements.

The Pre-Skyline Processing Agent (PSPA) prepares the results extracted from the database by the CSRSA for the running of the Skyline operator. The Cloud services returned and their dimensions are stored as tuples. The dimensions used are the user's requirements that are not "fixed", and thus are to be optimized, such as price (to be minimized), bandwidth (to be maximized), network latency (to be minimized)…

TABLE I. EXAMPLE OF FIXED REQUIREMENTS

Requirement	Value
Provider	Microsoft IBM Amazon…;
Service Model	IaaS PaaS SaaS
OS Serie	Windows Mac Unix…
OS Distribution	Windows XP Windows Vista Windows 7 Linux…
CPU Manufacturer	Intel IBM AMD…
CPU Gamme	Pentium Intel 64…
Industry	General Education Healthcare…
Category	General CRM E-procurement…

TABLE II. VALUE RANGE OF THE DIMENSIONS USED IN THE SKYLINE

Dimension	Value range
Storage space	0.14 – 4 000
Memory	128 – 16 000
Bandwidth	0 – 10
Latency	0 – 10 000
Cost	1 – 2 000
CPU speed	50 – 3 060

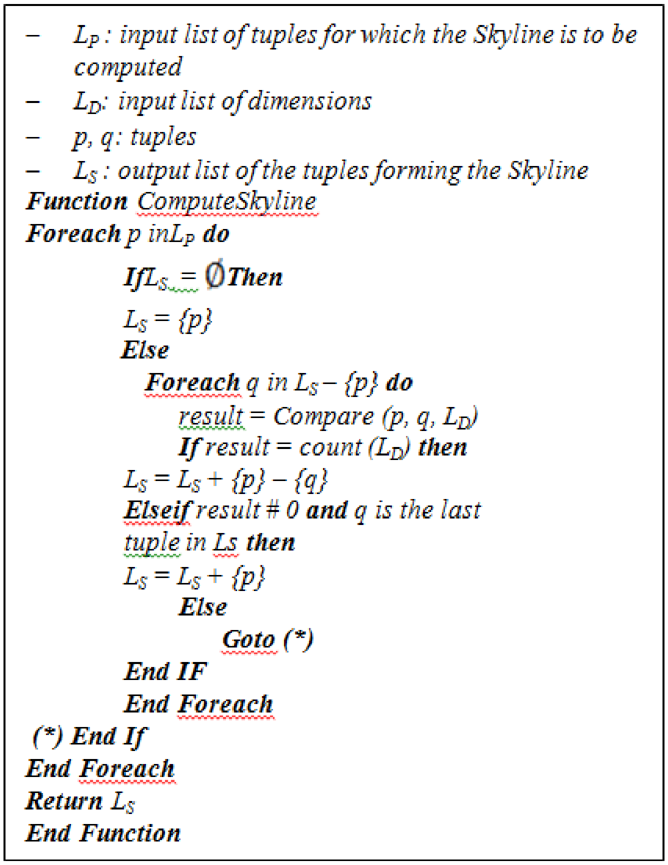
The CSRSA uses the Skyline [33], on the set of tuples returned by the PSPA, to determine which Cloud services are in the Skyline and meet the user's preferences.

The Skyline was introduced to meet the needs of users who want to select a set of points that optimize their requirements from a large set of data. Each point contained in the Skyline is not dominated by any other point, thus being better than all the points not contained in the Skyline for at least one criterion, and being equal to or better than them for all the other criteria. A criterion used by the Skyline is called dimension. For example, a user wants to rent a car at the minimum price with the maximum engine power. In this case, we have two dimensions upon which the selection is to be made: the first dimension is the price; the second is the engine power. The Skyline algorithm will compute the Skyline, which will contain all the cars that are not dominated by any other car. In other words, for each car returned in the Skyline, there is no car outside the Skyline that is less expensive and has more engine power at the same. Thus, a user will find their favorite car in the Skyline, no matter how they weight their preferences toward the dimensions.

In our case, using the Skyline allows the user to specify the criteria they want to optimize and to get the Cloud services that are not dominated by any other Cloud service, that is to say Cloud services for which there exists no better Cloud service for all the criteria specified.

There are two major ways to compute the Skyline. One is to extend existing database systems with the logical Skyline operator. The other is to use algorithms. Many algorithms may be used such as the Block-Nested Loops algorithm (BNL) [33], the Divide and Conquer algorithm (D&C) [56, 57], the B-Tree algorithm [58], etc. We used the BNL algorithm (*Fig.* 2) because it is efficient, simple to implement. It has a complexity of O(n^2) in the worst cases and O(n) in the best.

As presented in [36], the BNL algorithm consists of comparing tuples among them to determine the ones that are not dominated by any other. It is done by keeping dominating tuples in the main memory and by comparing each new tuple to them. In each iteration, a new tuple is read from the input list of tuples. If the new tuple is dominated by one of the existing tuples in the main memory, it is eliminated. If it dominates a tuple in the main memory, the dominated tuple is eliminated, and the new tuple is added to the main memory to be compared to future tuples. If the new tuple is incomparable, which means that it is neither dominated by nor dominating any tuple in the main memory, it is added to the main memory. At the end of all iterations, only tuples that are not dominated by any other tuple are kept in the main memory. These tuples form the Skyline. The function Compare (p, q, L_D) (*Fig. 2*) as presented in [36] compares the tuples p and q in all the dimensions in the list L_D. The result returned varies between 0 (when q dominates p) and the number of dimensions n (when p dominates q). Any other result in this range means that p and q are not comparable. All the tuples contained in the output list are incomparable among them. This means, if we take any given two tuples, each one would be at least better than the other in some dimensions, and at least worse in others.



```
– $L_P$ : input list of tuples for which the Skyline is to be computed
– $L_D$: input list of dimensions
– p, q: tuples
– $L_S$ : output list of the tuples forming the Skyline
Function ComputeSkyline
Foreach p in$L_P$ do
        If$L_S$ = ∅Then
        $L_S$ = {p}
        Else
          Foreach q in $L_S$ – {p} do
             result = Compare (p, q, $L_D$)
             If result = count ($L_D$) then
        $L_S$ = $L_S$ + {p} – {q}
        Elseif result # 0 and q is the last
        tuple in Ls then
        $L_S$ = $L_S$ + {p}
             Else
                 Goto (*)
        End IF
        End Foreach
 (*) End If
End Foreach
Return $L_S$
End Function
```

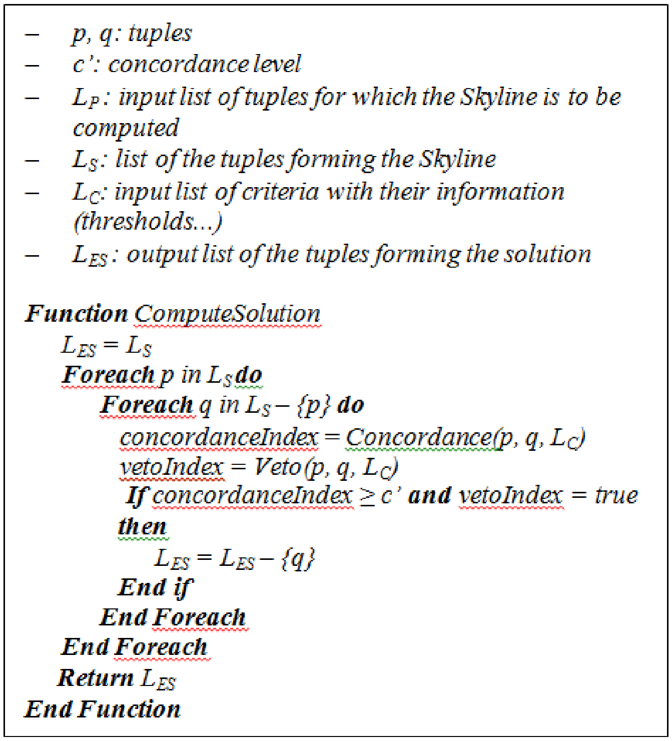
Algorithm 1: Algorithm of the Skyline Agent as presented in [35]

The Skyline doesn't allow arbitrating between incomparable tuples. This comes from the fact that all dimensions are considered to have the same importance, which is not always true to users. To overcome this limitation, we thought of using outranking methods, more specifically ELECTRE methods.

As seen previously, the Cloud Service Research and Selection Agent (noted CSRSA) uses the Skyline on the set of tuples returned by the Pre-Skyline Processing Agent (noted PSPA) to determine which Cloud services are in the Skyline and meet the user's preferences. To do so, the CSRSA uses the BNL algorithm as showed in *Fig. 2*. In order to refine the results returned, we adjusted our prototype (*Fig. 3*) by adding an ELECTRE IS agent. This agent uses the algorithm presented in *Fig. 4* to apply the ELECTRE IS to the Skyline list returned by the CSRSA.

The Skyline list becomes the input list of alternatives on which the ELECTRE IS method is applied. Thus, a pairwise comparison is made and the concordance and veto indexes are determined. If the validating condition is verified, the alternative that is outranked by the other alternative is deleted from the list of the final solution. Thus, the output list contains only the alternatives that are incomparable both to the Skyline and the ELECTRE IS agents.

In the next section we present Cloud related Quality of Service (QoS) requirements and in particular the ones we use as new dimensions/criteria in the CSRS System.



- *p, q: tuples*
- *c': concordance level*
- *L_P: input list of tuples for which the Skyline is to be computed*
- *L_S: list of the tuples forming the Skyline*
- *L_C: input list of criteria with their information (thresholds…)*
- *L_{ES}: output list of the tuples forming the solution*

```
Function ComputeSolution
    $L_{ES} = L_S$
    Foreach p in $L_S$ do
        Foreach q in $L_S - \{p\}$ do
            concordanceIndex = Concordance(p, q, $L_C$)
            vetoIndex = Veto(p, q, $L_C$)
            If concordanceIndex $\geq$ c' and vetoIndex = true then
                $L_{ES} = L_{ES} - \{q\}$
            End if
        End Foreach
    End Foreach
    Return $L_{ES}$
End Function
```

Algorithm 2: Algorithm of the ELECTRE IS Agent as presented in [36]

IV. Quality of Service

As users increasingly turn to Cloud services providers to purchase their services, they are more and more demanding when it comes to Quality of Service (QoS).

QoS is defined as being the *"totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service"* [59].

There are many important QoS parameters to take into account when looking for a Cloud service, such as time, cost, reliability, security [60]…

The requirements of Cloud users regarding QoS parameters are described in Service-Level Agreements (SLAs) to help providers manage the services contracted and maintain the overall level of quality agreed on with end-users [61]. And since Cloud resources are consumed simultaneously by multiple users/tenants, providers have to dynamically allocate Cloud resources among them while guaranteeing the QoS level agreed on for every one of them. So for Cloud users, QoS and SLA are key factors when they select Cloud services.

Measuring the performance of Cloud services is not an obvious task. For one part, there is the question of quantifying parameters that are essentially qualitative. Many works have tried to provide a set of QoS parameters that can be used by Cloud users to select the most adapted services.

In [62], Cao et al. present a QoS-assured Cloud Computing architecture to answer QoS-related requests from users. This architecture consists of X layers: physical device and virtual resource, cloud service provision, cloud service management, and multi-agent. The QoS attributes considered are related both to the users and the providers of cloud services. Many attributes were defined, namely response time, cost, availability, reliability and reputation.

In [63], Ferreti et al. propose a middleware architecture to configure, manage and optimize cloud services in accordance with users' QoS requirements such as timeliness, scalability, availability, and security. The proposed architecture integrates three main components, namely dynamic resources configuration, platform monitoring, dispatching and load balancing of requests and resources. The cloud computing environment resulting is labeled "QoS-aware".

Bouguettaya et al. presented in [64] a QoS-based approach for the selection of cloud services for composition purposes. The aim of this approach is to compose a cloud service that answers the QoS requirements of end-users from multiple composite services provided by different providers. It is done by constructing the composition schema based on the user's request, then selecting the optimal composition plan based on the end-user's QoS requirements. The QoS attributes used are throughput, response time, and cost.

In [65], Zheng et al. presented CloudRank, a QoS ranking prediction framework for cloud services that takes into account users' experiences. In this work, QoS attributes are divided in two categories: client-side and server-side. The latter include response time, throughput, failure probability, etc. and are the ones used in CloudRank. It also uses similarity to determine the degree to which the current user is similar to other users in order to predict which Cloud services would be more interesting.

In [66], Nathuji et al. developed Q-Clouds, a control framework that supports QoS-aware cloud environments, as presented in [63]. Q-Clouds adapts the allocation of resources to absorb the effect of performance interferences that are bound to happen, since many users share the same resources. This is done while taking into account the QoS requirements of users.

In [67], Serrano et al. address the challenge of QoS and SLAs management in Cloud environment by defining a new Cloud model called SLA Aware Service (SLAaaS) and a new language to describe QoS-oriented cloud SLAs, called CSLA, that is inspired from WSLA (SLA for web services) [68] and SLA for Service Oriented Architecture (SLA@SOI) [69].

CSLA formalizes the SLA between users and providers by translating QoS requirements into clauses combined using Boolean operators. QoS attributes adopted in this work are related to performance, availability, reliability and cost.

Another prominent work is the Service Measurement Index (SMI) developed by the Cloud Service Measurement Initiative Consortium (CSMIC) [49].

Service Measurement Index (SMI)is defined by the CSMIC as being *"a set of business-relevant Key Performance Indicators (KPI's) that provide a standardized method for measuring and comparing a business service regardless of whether that service is internally provided or sourced from an outside company"*.

SMI measures performance using six categories (*Fig. 4*): agility, risk, security, cost, quality, and capability. Each category contains many attributes.

- Agility: one of the main reasons why organizations choose to move to the Cloud is to increase their agility in order to quickly adapt to their ever-changing business environment. In order to measure agility, SMI proposed a set of parameters that show how quickly and efficiently providers integrate new capabilities to answer users' evolving needs. We retained from these parameters portability;

- Risk: risk is an inherent in IT, the main objective of organizations being not to annihilate it, but to minimize its causes and effects. The main issue for users when choosing a Cloud service provider is verifying the reputation of the latter and making sure they have an efficient risk management strategy. We chose to retain as a parameter the number of risk management certifications the provider has;

- Security: moving to the Cloud can be challenging for users when it comes to entrusting their critical data to providers. Many aspects must be addressed, mainly privacy and data loss. This issue is also related to the laws governing the geographical location of data storage. We chose to initially retain data loss as a parameter;

- Cost: another main reason why organizations and users move to the Cloud is to optimize their IT-related costs. With the Cloud being a pay-as-you-go utility, users are guaranteed to pay solely for resources they consume, without surplus or need to invest or manage the underlying infrastructure. We chose to retain the two cost parameters defined by SMI, namely the on-going cost, which is the cost regularly paid by the Cloud tenant in exchange for the resources they use, and the acquisition cost, which is the cost of the changes necessary to move to the Cloud;

- Quality: when choosing to move to the Cloud, users usually worry about the quality of the services offered by providers, especially as regards their reliability and availability. This comes from the fact that users will be moving their data out of their control and into that of the provider. We chose to retain as parameters availability and service response time;

- Capability: SMI proposes to measure the overall ability of a cloud services provider to satisfy users' requirements by comparing the services offered to standards. We refrained from using a parameter to translate this characteristic since, in our knowledge; there are no unified Cloud standards yet.

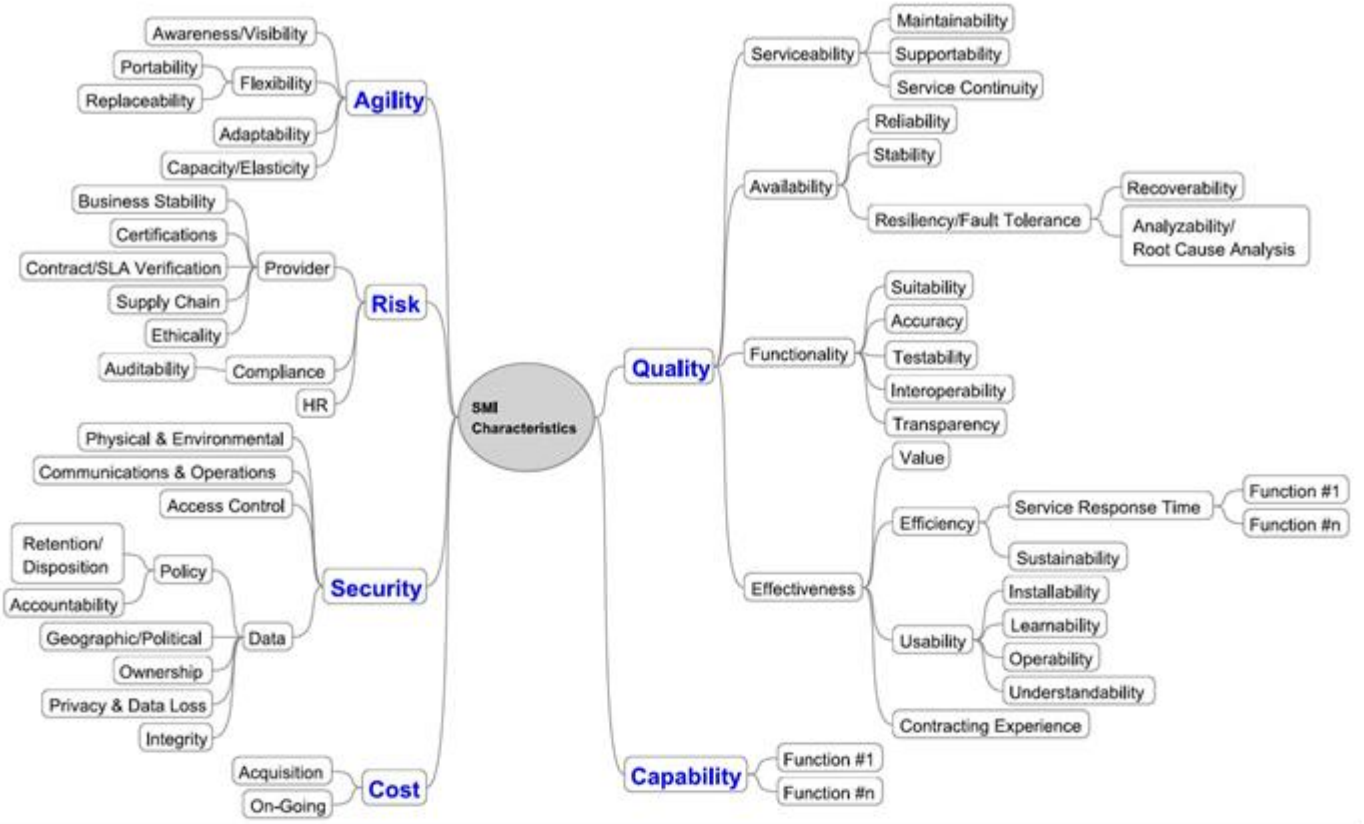


Fig. 2. Service Measurement Index (SMI) characteristics [49]

As explained above, we have chosen to use one or two parameters in each category. These parameters will be used as additional dimensions for the Skyline and criteria for ELECTRE IS. For each new dimension/criterion, we propose a measurement method.

TABLE III. QOS PARAMETERS USED AS DIMENSIONS FOR THE SKYLINE

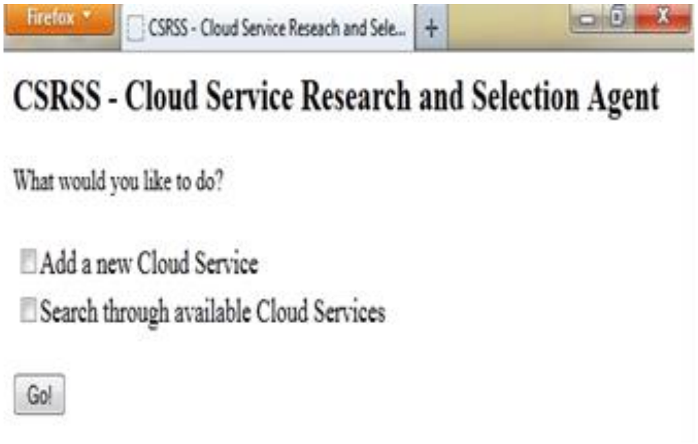
Category	Dimension/Criterion	Detail	Comparison sense
Agility	Portability	$\frac{\text{number of OS compatible with the service}}{\text{number of OS required by the user}}$	Maximize
Risk	Number of risk management certifications obtained by the provider	–	Maximize
Security	Data Loss	Number of data loss related incidents	Minimize
Cost	Acquisition cost	–	Minimize
	On-going cost	–	Minimize
Quality	Service response time	$\frac{\text{average response time (ms)}}{\text{maximum Sresponse time defined in the SLA (ms)}}$	Minimize
	Availability	$\frac{\text{time during which the service is unavailable (ms)}}{\text{total time of use (ms)}}$	Minimize

In the next section we present the implementation of the algorithm and its performance along with some screenshots illustrating its execution.

V. EXPERIMENTATION AND RESULTS

The platform we used for the experiments [35] is an HP workstation with a 3.30 GHz processor, 4 GB of main memory, Windows Server 2008 as operating system and MS SQL Server 2008 as DBMS. The algorithm is implemented using ASP.net to obtain a web-based system that can be accessed from any web client anytime the user is connected to the Internet.

The CSRSS start page (*Fig. 3*) allows the user to either add a new Cloud service to the database or search for Cloud services that match their requirements.



CSRSS - Cloud Service Research and Selection Agent

What would you like to do?

□ Add a new Cloud Service
□ Search through available Cloud Services

Go!

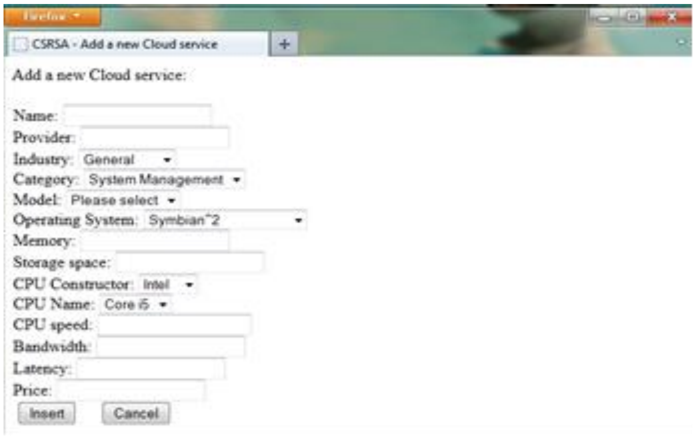
Fig. 3. The CSRA start page

If the user chooses to add a new Cloud Service, they are taken to another page (*Fig. 4*) where they first enter the name of the Cloud service in question so a search can be made to make sure that it doesn't already exist in the database. Afterwards, the user enters the different information such as the Cloud service's provider, model (IaaS, PaaS or SaaS), industry, memory, price…

If the user checks the second option (Search through available Cloud Services), they are taken to the CSRSS page (*Fig. 5*) that allows to make an advanced search through the database and to run the algorithm on the returned result in order to obtain the final refined set of Cloud services.

The user can fill out one or many information about the Cloud service(s) they are searching for. For information such as price, memory, storage space, bandwidth... they can either give a specific value or specify that they are the dimensions to be used when computing the Skyline. For each dimension, the user specifies if it is to be minimized or maximized. They also specify its importance (on a scale from "not important" to "extremely important") that is to be translated into a weight in order to use the ELECTRE IS method.

We worked on the same generated data as in [35]. This data consists of 50 000 Cloud services which we completed with the new 7 dimensions described in Table III. We also chose to disregard dimensions that are too oriented for a specific Cloud service model, such as RAM and CPU speed, which are mostly relevant in IaaS environments, for instance. Thus, we obtained a total 10 dimensions/criteria for each cloud service. Dimensions' values are randomly generated within the ranges specified in Table IV hereafter.

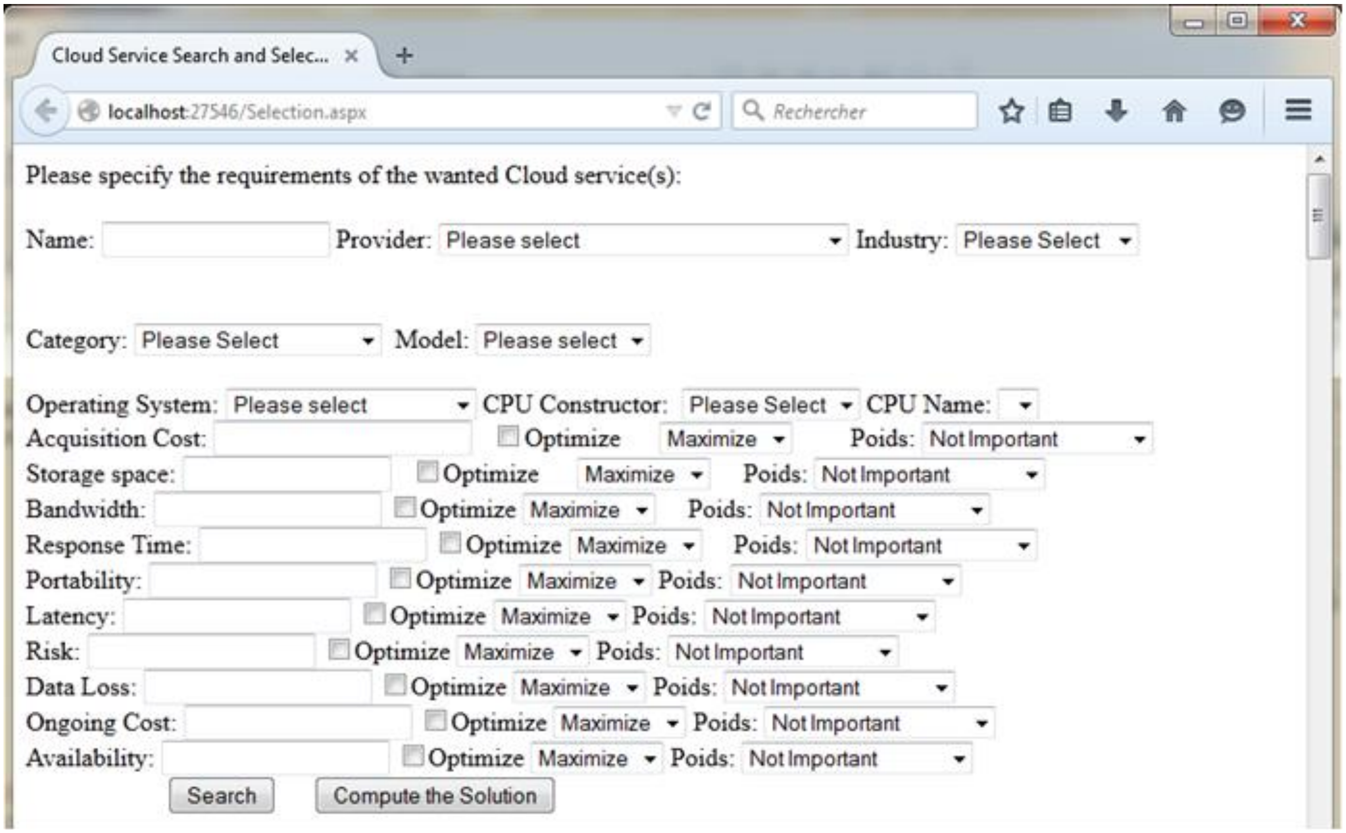


Add a new Cloud service:

Name:
Provider:
Industry: General
Category: System Management
Model: Please select
Operating System: Symbian^2
Memory:
Storage space:
CPU Constructor: Intel
CPU Name: Core i5
CPU speed:
Bandwidth:
Latency:
Price:

Insert Cancel

Fig. 4. The CSRA page to add a new Cloud service


Cloud Service Search and Selec...
localhost:27546/Selection.aspx
Rechercher
Search
Compute the Solution

Please specify the requirements of the wanted Cloud service(s):

Name: Provider: Please select Industry: Please Select

Category: Please Select Model: Please select

Operating System: Please select CPU Constructor: Please Select CPU Name:
Acquisition Cost: □ Optimize Maximize Poids: Not Important
Storage space: □ Optimize Maximize Poids: Not Important
Bandwidth: □ Optimize Maximize Poids: Not Important
Response Time: □ Optimize Maximize Poids: Not Important
Portability: □ Optimize Maximize Poids: Not Important
Latency: □ Optimize Maximize Poids: Not Important
Risk: □ Optimize Maximize Poids: Not Important
Data Loss: □ Optimize Maximize Poids: Not Important
Ongoing Cost: □ Optimize Maximize Poids: Not Important
Availability: □ Optimize Maximize Poids: Not Important

Fig. 5. The CSRA search and/or computation of the Solution page

TABLE IV. VALUE RANGE OF THE DIMENSIONS USED IN THE SKYLINE

Dimension	Value range	Comparison sense
Storage space	0.14 – 4 000	Maximize
Bandwidth	0 – 10	Maximize
Latency	0 – 10000	Minimize
Portability	0.03 – 400	Maximize
Risk	0 – 400	Maximize
Data Loss	0 – 9 000	Minimize
Acquisition cost	1 – 20 000	Minimize
On-going cost	0.1 – 2000	Minimize
Service response time	0 – 40	Minimize
Availability	0 – 1 000	Minimize

We executed our program varying the number of dimensions from 1 to10 and the input size from 100 to 50 000 cloud services (*Fig. 6*).

We also compared the results obtained using ELECTREIsSkyline algorithm with those of using the Skyline algorithm as presented in [35]. We did so for an input list consisting of 50 000 Cloud services and varying the number of dimensions from 1 to 10 (Table V and *Fig. 7*).

TABLE V. tHE SIZE OF THE FINAL SOLUTION DEPENDING ON THE NUMBER OF CRITERIA AND THE ALGORITHM USED FOR 50 000 CLOUD SERVICES

Number of criteria	Final solution's size	
	Skyline algorithm	ELECTREIsSkyline algorithm
1	18	1
2	25	7
3	1436	205
4	3111	539
5	3448	957
6	4316	1 546
7	6918	3 187
8	5285	3 688
9	5286	3 945
10	7 360	3 610

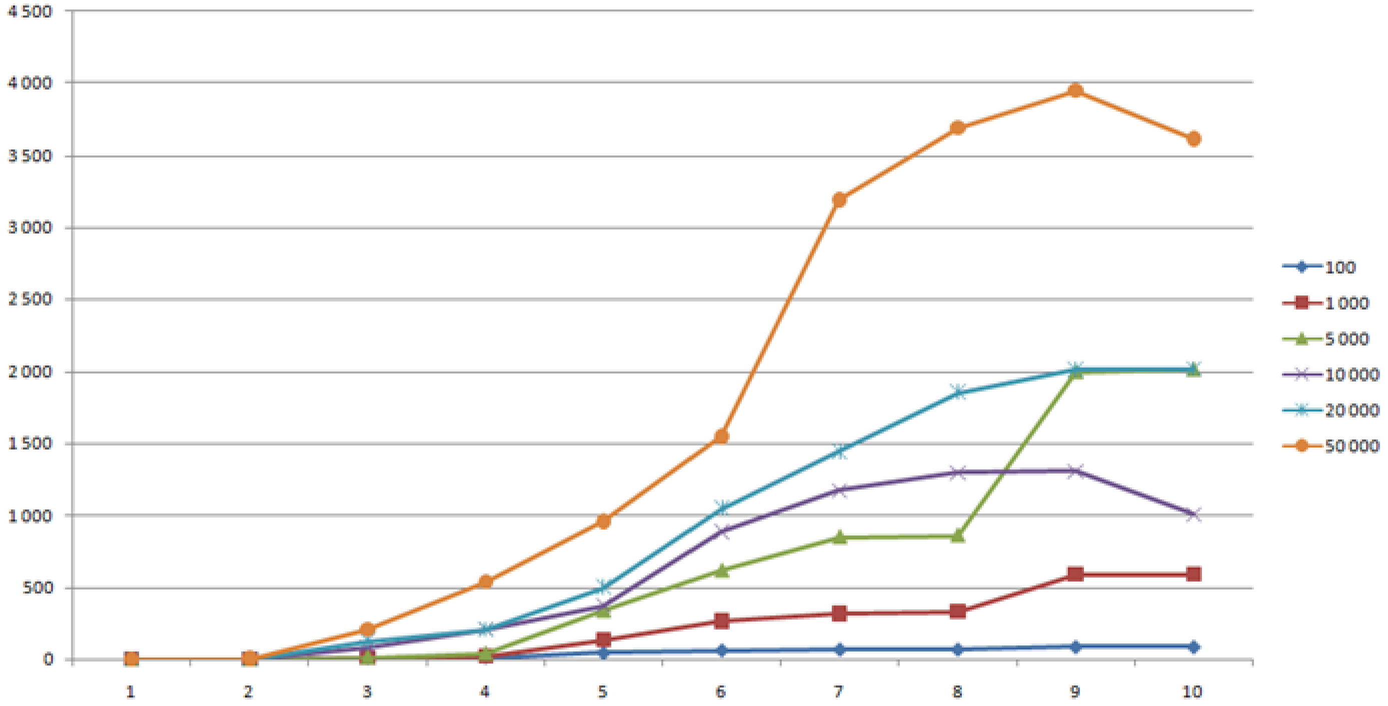


Fig. 6. The size of the solution depending on the number of dimensions and the input size for the ELECTREIsSkyline Algorithm

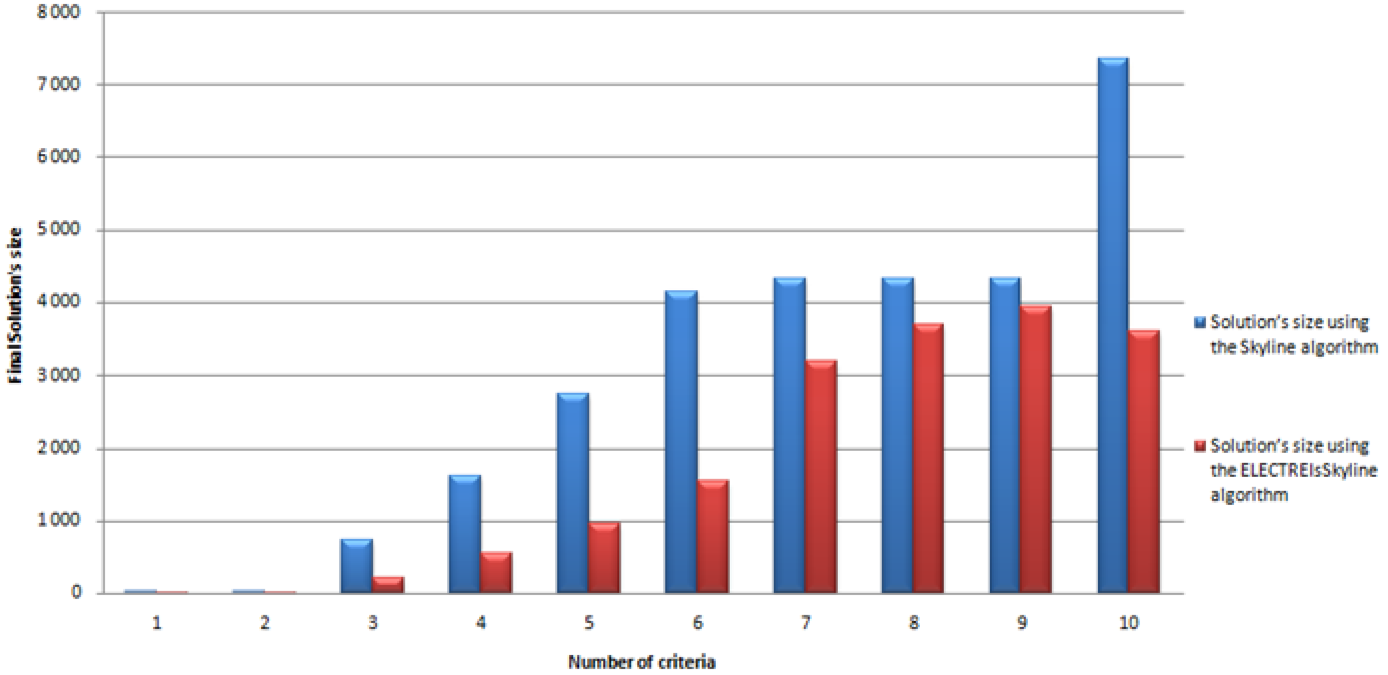


Fig. 7. The size of the solution depending on the number of dimensions for each used algorithm

The use of the ELECTREIsSkyline algorithm proves to be more efficient in determining the cloud services that best match the users' requirements, including QoS requirements. It is due to the fact that users' preferences towards each criterion are taken into account. Thus, dimensions that had the same weight to the Skyline algorithm have different weights in ELECTREIsSkyline, depending on their importance in the decision making process. So, cloud services that were incomparable when using the Skyline become comparable when using ELECTREIsSkyline. The size of the final solution can be reduced to contain only 6% of the input size, returning a total of 3 000 Cloud services from the 50 000 in the input list, while taking into consideration all 10 criteria and their respective weights. However, the final result is highly dependent on the values of the dimensions/criteria in the input size, and the weights attributed by users to these criteria. Also, it is common that when augmenting the input size, one or more new cloud services prove to be highly efficient and allow eliminating many others, reducing significantly the size of the output. These cloud services are called “killer tuples”. On the other hand, it is also common that new cloud services turn out to be incomparable with those contained in the solution list, which contributes to augmenting its size. Thus, the final solution’s size is highly depending on the quality of input data and the user’s preferences.

VI. Conclusion

With many Cloud providers offering their services, Cloud users may be at loss when wanting to choose an adapted cloud service to their needs. To address this issue, we have presented in [35] and [36] the Cloud Service Research and Selection System (CSRSS) that allows users to select cloud services that best suit them by specifying the requirements they are looking for. In this work, we tried to ameliorate the CSRSS by addressing the issue of QoS and adapting it to integrate QoS parameters, giving users the possibility to specify the values of the QoS attributes they require.

References

[1] S. Casselman, “Virtual computing and the virtual computer”, FPGAs for Custom Computing Machines, Proceedings IEEE Workshop, 1993

[2] P.H ENSLOW, “What is a "distributed" data processing system?”, Computer, Vol. 11, No 1, 1978

[3] I. Foster, Y. Zhao, I. Raicu and S. Lu, “Cloud Computing and grid computing 360-degree compared”, Grid Computing Environments Workshop, GCE'08, November 2008

[4] D. Roman et al., “Web service modeling ontology”, Applied ontology, Vol. 1, No 1, 2005

[5] M. BELL, “Introduction to Service-oriented Modeling”, Service-oriented Modeling: Service Analysis, Design, and Architecture. Wiley & Sons, Vol. 3, 2008

[6] Wang, Lizhe, et al. “Scientific Cloud Computing: Early Definition and Experience”, HPCC, Vol. 8, 2008

[7] A. Vouk, “Cloud Computing–issues, research and implementations”, Journal of Computing and Information Technology, Vol. 16, No 4, 20

[8] R. Buyya, C.S. Yeo and S. Venugopal, “Market-oriented Cloud Computing: Vision, hype, and reality for delivering it services as computing utilities”, High Performance Computing and Communications, 2008, HPCC'08, 10th IEEE International Conference, Ieee, 2008

[9] P. Mell and T. Grance, “The NIST definition of Cloud Computing”, NIST special publication, 2011

[10] A.Fox, G. Rean, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, “Above the clouds: A Berkeley view of Cloud Computing”, Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS 28, 2009

[11] L. Vaquero, L. Rodero-Merino, J. Caceres and M. Lindner, “A Break in the Clouds: Towards a Cloud Definition”, ACM SIGCOMM Computer Communication Review, Vol. 39, Number 1, January 2009

[12] Google Drive, https://drive.google.com

[13] Salesforce, http://www.salesforce.com

[14] D. Cheng, “PaaS-onomics: A CIO’s Guide to using Platform-as-a-Service to Lower Costs of Application Initiatives While Improving the Business Value of IT”, Tech. rep., LongJump, 2008

[15] Force, http://www.force.com

[16] Google App Engine, https://appengine.google.com

[17] Windows Azure, http://www.windowsazure.com

[18] L. Karadsheh, “Applying security policies and service level agreement to IaaS service model to enhance security and transition”, Computers & Security, Vol. 31, Issue 3, May 2012, pp. 315-326

[19] S. Radack, “Cloud Computing: A Review of Features, Benefits, and Risks, and Recommendations for Secure, Efficient Implementations”, NIST, ITL Bulletin, June 2012

[20] Amazon, http://aws.amazon.com/fr/ec2

[21] Microsoft SQL Azure, http://www.windowsazure.com

[22] L. Youseff, L. Butrico and M. Da Silva, “Toward a Unified Ontology of Cloud Computing”, Grid Computing Environments Workshop, November 2008

[23] P. Costa et al. "NaaS: Network-as-a-Service in the Cloud" Proceedings of the 2nd USENIX conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services, Vol. 12, 2012

[24] E. B. Dudin and G. Yu Smetanin, “A review of Cloud Computing”, Scientific and Technical Information Processing, Vol. 38, No 4, 2011

[25] M. Christodorescu, R. Sailer, D. L. Schales, D. Sgandurra and D. Zamboni, “Cloud security is not (just) virtualization security: a short paper”, Proceedings of the 2009 ACM workshop on Cloud Computing security, ACM, 2009

[26] R. Accorsi, “Business process as a service: Chances for remote auditing”, Computer Software and Applications Conference Workshops (COMPSACW), 2011 IEEE 35th Annual, IEEE, 2011

[27] M. Zhou, R. Zhang, D. Zeng and W. Qian, “Services in the Cloud Computing era: A survey”, Universal Communication Symposium (IUCS), 2010 4th International IEEE, 2010

[28] J. Wu, L. Ping, X. Ge, Y. Wang, and J. Fu, “Cloud storage as the infrastructure of Cloud Computing”, Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference IEEE, 2010

[29] M. Brock and A. Goscinski, “Toward Ease of Discovery, Selection and Use of Clusters within a Cloud,” 2010 IEEE 3rd International Conference on Cloud Computing, July 2010

[30] S. Rao, N. Rao and E. Kusuma Kumari, “Cloud Computing: An Overview”, Journal of Theoretical and Applied Information Technology, Vol. 9, No. 1, November 2009

[31] K. Sims, “IBM Blue Cloud Initiative Advances Enterprise Cloud Computing”, 2009

[32] K. Jeffery and B. Neidecker-Lutz, “The future of Cloud Computing”, European Commission, Information Society and Media, 2010

[33] S. Börzsönyi, D. Kossmann, and K. Stocker, “The Skyline operator”, International Conference on Data Engineering (ICDE), 2001

[34] B. Roy, “The outranking approach and the foundation of the ELECTRE methods”, Theory and decision, Vol. 31, Issue 1, 1991

[35] A. Idrissi and M. Abourezq, “Skyline in Cloud Computing”, Journal of Theoretical and Applied Information Technology, Vol. 60, No. 3, February 2014

[36] A. Idrissi and M. Abourezq, “Introduction of an outranking method in the Cloud computing research and Selection System based on the Skyline”, Proceedings of the International Conference on Research Challenges in Information Science (RCIS), 2014

[37] W. Zeng, Y. Zhao and J. Zeng, “Cloud Service and Service Selection Algorithm Research”, of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation

[38] J. Kang and K. M. Sim, “A Cloud Portal with a Cloud Service Search Engine”, International Conference on Information and Intelligent Computing IPCSIT, Vol.18, 2011

[39] P. Resnik, “Semantic similarity in a taxonomy: an information-based measure and its application to problem of ambiguity in natural language”, Journal of Artificial Intelligence Research, Vol. 11, 1999

[40] J. Kang and K. M. Sim, “Cloudle : An Agent-based Cloud Search Engine that Consults a Cloud Ontology”, Cloud Computing and Virtualization Conference, 2010

[41] T. Han and K. M. Sim, “An Ontology-enhanced Cloud Service Discovery System”, IMECS 2010 Vol. 1, March 17 – 19 2010, Hong Kong

[42] H. Yoo, C. Hur, S. Kim, and Y. Kim, “An Ontology-based Resource Selection Service on Science Cloud”, International Journal of Grid and Distributed Computing, Vol. 2, No. 4, December 2009

[43] C. Zeng, X. Guo, W. Ou and D. Han, “Cloud Computing Service Composition and Search Based on Semantic”, Cloud Computing, Vol. 5931, 2009, pp. 290-300

[44] L. Sun, H. Dong, F. K. Hussain, O. K. Hussain, and E. Chang, "Cloud service selection: State-of-the-art and future research directions." Journal of Network and Computer Applications 45, 2014

[45] T.L. Saaty, “The Analytic Hierarchy Process for Decision in a Complex World”, Pittsburgh, PA: RWS Publications, 1980

[46] T. L. Saaty, “Decisions with the analytic network process (ANP)” University of Pittsburgh (USA), ISAHP 96 (1996)

[47] C. W. Churchman, R. L. Ackoff and E.L. Arnoff, “Introduction to Operations Research”, New York: Wiley, 1957

[48] S. K. Garg, S. Versteeg and R. Buyyaa, “SMICloud: A framework for ranking of Cloud Computing services”, IEEE International Conference on Utility and Cloud Computing, 2011

[49] Cloud Service Measurement Index Consortium (CSMIC), SMI framework, http//cloudcommons.com/servicemeasurementindex

[50] M. Godse and S. Mulik, “An approach for selecting software-as-a-service (SaaS) product”, Proceedings of the IEEE international conference on Cloud Computing (CLOUD), Bangalore, 2009

[51] R. Karim, C. Ding and A. Miri, “An end-to-end QoS mapping approach for Cloud service selection”, Proceedings of the IEEE 9th world congress on services (SER-VICES), Santa Clara Marriott, CA, 2013

[52] M. Menzel, M. Schönherr and S. Tai, “$(MC2)^2$: criteria, requirements and a software prototype for Cloud infrastructure decisions”, Softw Pract Exp November 2013, Vol. 43, Issue 11

[53] N. Limam and R. Boutaba, “Assessing software service quality and trustworthiness at selection time”, IEEE Trans Softw Eng 2010, Vol. 36, Issue 4

[54] S. Silas, E. B. Rajsingh and K. Ezra, “Efficient service selection middleware using ELECTRE methodology for Cloud environments”, Information Technology Journal, Vol. 11, Issue 7

[55] D. Androcec, N. Vrcek and J. Seva, “Cloud Computing Ontologies: A Systematic Review”, MOPAS 2012, The Third International Conference on Models and Ontology-based Design of Protocols, Architectures and Services, 2012

[56] H. Kung, F. Luccio and F. Preparata, “On finding the maxima of a set of vectors”, Journal of the ACM, Vol. 22, Issue 4, October 1975

[57] F. Preparata and M. Shamos, “Computational Geometry: An Introduction”, Springer-Verlag, New York, 1985

[58] D. Comer, “The Ubiquitous B-Tree”, ACM Computing Surveys, Volume 11, June 1979

[59] "Terms and definitions related to quality of service and network performance including dependability", International Telecommunication Union Recommendation, 1994

[60] R. Buyya, C. S. Yeo and S. Venugopal, “Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities”, High Performance Computing and Communications, 2008, HPCC'08, 10th IEEE International Conference

[61] M. E. Shacklett, “Five Key Points for Every SLA”, http://content.dell.com/us/en/enterprise/d/large-business/key-points-for-sla.aspx, 2011

[62] B. Q. Cao, B. Li and Q. M. Xia, “A Service-Oriented QoS-Assured and Multi-AgentCloud Computing Architecture”, Cloud Computing, Springer Berlin Heidelberg

[63] S. Ferretti, V. Ghini, F. Panzieri, M. Pellegrini and E. Turrini, “QoS–aware Clouds”, Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference

[64] Z. Ye, A. Bouguettaya and X. Zhou, “QoS-Aware Cloud Service Composition Based on Economic Models”, Service-Oriented Computing, Springer Berlin Heidelberg

[65] Z. Zheng, X. Wu, Y. Zhang, M. R. Lyu and J. Wang, “QoS Ranking Prediction for Cloud Services”, Parallel and Distributed Systems, IEEE Transactions, Vol. 24, Issue 6

[66] R. Nathuji, A. Kansal and A. Ghaffarkhah, “Q-Clouds: Managing Performance Interference Effects for QoS-Aware Clouds”, Proceedings of the 5th European conference on Computer systems, ACM, 2010

[67] D. Serrano, S. Bouchenak, Y. Kouki, T. Ledoux, J. Lejeune, J. Sopena and P. Sens, “Towards QoS-Oriented SLA Guarantees for Online Cloud Services”, Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium

[68] H. Ludwig, A. Keller, A. Dan, R. P. King, and R. Franck, “Web Service Level Agreement (WSLA) Language Specification,” IBM, Tech. Rep., 2003

[69] “Sla@soi,” sla-at-soi.eu/, 2012.

Cyberspace Forensics Readiness and Security Awareness Model

Aadil Al-Mahrouqi
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland

Sameh Abdalla
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland

Tahar Kechadi
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland

Abstract—**The goal of reaching a high level of security in wire- less and wired communication networks is continuously proving difficult to achieve. The speed at which both keepers and violators of secure networks are evolving is relatively close. Nowadays, network infrastructures contain a large number of event logs captured by Firewalls and Domain Controllers (DCs). However, these logs are increasingly becoming an obstacle for network administrators in analyzing networks for malicious activities. Forensic investigators mission to detect malicious activities and reconstruct incident scenarios is extremely complex considering the number, as well as the quality of these event logs. This paper presents the building blocks for a model for automated network readiness and awareness. The idea for this model is to utilize the current network security outputs to construct forensically comprehensive evidence. The proposed model covers the three vital phases of the cybercrime management chain, which are:**

1) Forensics Readiness, 2) Active Forensics, and 3) Forensics Awareness.

Keywords—Network Forensics; Forensics Readiness; Network Security; Active Forensics; Reactive Forensics; Forensics Awareness and Network Security model

I. Introduction

The cybercrime landscape has increased dramatically with the use of more sophisticated techniques and greater knowledge of cybercrime. There are many challenges faced by todays digital forensics. The lack of both funding and qualified professionals, as well as cross-jurisdictional legal struggles are just a sample of the main body of issues [1].The first Digital Forensics Research Workshop (DFRWS) [2] was held in Utica, N.Y., in 2001. DFRWS provided the first proper framework and presented guidelines for conducting a technical digital investigation.

It is now evident that the cyber-infrastructure requirements and associated data management systems are becoming large in number, highly dynamic in nature, and exceptionally attractive for cyber-crime activities [3]. Protecting the sensitive data cyber-infrastructure portals are relying on information security for daily activities which is not a trivial task. The techniques used to perform cybercrimes are becoming relatively sophisticated with the firewalls protecting them. Reaching high-levels of data protection in both wired and wireless networks, in order to face recent cybercrime approaches is a challenge that is continuously proving hard to achieve.

Since that first workshop, many scholars have worked to make digital evidence easier to demonstrate by establishing many types of graphs in order to represent evidence and attack scenarios. The scholars utilized a mathematical formula and algorithms to construct these graphs to recognize the patterns of the attack [4]. Unfortunately, most of these graphs provide a high-level, abstract view of the complex attack [5]. Examples of investigation graphs consist primarily of scenario graphs, forensics graphs, logic exploitation graphs, attack graphs, and evidence graphs [6]. The digital systems can be described mathematically as a finite state machine and can represents this information in the form of a graph (nodes and arrows) [7].

Figure 1 shows the cyber-crime management chain, it consists of four stages namely; proactive (readiness), active, reactive and awareness. The first phase in the cyber-crime chain is the proactive phase and its goal is to prepare the target network to automatically prevent and detect the attack or illegal activities before the network gets infected, such as user authentication and system capable of avoiding programming errors and information protection e.g. Privacy Preserving Data Mining (PPDM). The active approach in the cyber-crime chain is used to detect and analyse anomaly activities and attack in real-time e.g. Firewalls. The reactive approach deals with the analysis of the victim network or assesses the incident after it happens e.g. Host-based (HIDs) and Network-based (NIDs) Intrusion detection system. Finally, the awareness approach deals with the training and awareness proposal. These works take into consideration the important factors during forensics investigation, for example; cost, time, low incident impacts, facilities network investigation procedures, high quality outcomes, organization reputation and business activities disruption. Furthermore, the aim to propose an attack and evidence integration graph is to increase the efficiency of investigation results. In addition, the data flow in the proposed model is designed based on the network OSI model. On the other hand, this paper presents a forensics awareness model designed to generate a best practice for system administrations and forensics investigators to learn about security vulnerabilities from previous cases in the network infrastructure, as well as different sources.

The remaining part of this paper is organized as follows; Section 2 outlines previous work. Section 3 describes the proposal model. Section 4 establishes a case study with the aims to give an idea of how to create a criminal graph. Finally, section 5 deals with the conclusion and some perspectives on future work.

II. Previous Work

The authors [8] performed an in-depth survey for events admissibility in the Irish court of law. Overall, the legal review is mainly focused on different primary areas: the admissibility and authentication of digital evidence and focuses mainly on Irish law. Admissibility refers to a set of lawful tests carried out by a judge for forensic assessment of the finding evidence. Trustworthy means that an accurate copy of digital evidence was acquired, and that it has continued to be unchanged since it was recovered. Authentication is a process to check the reliability of digital evidence. The judge summarizes five issues that must be considered when evaluating whether evidence will be admitted, namely; not unduly prejudicial, best evidence, not hearsay or admissible hearsay, authenticity and relevance.

Wang & Daniels [9], in their proposed evidence graph model seek to facilitate the presentation and manipulation of intrusion evidence. This model aims to reduce the redundancy in firewall output intrusion alerts. The proposed architecture facilitates the evidence presentation process and provides automated intrusion evidence analysis. The evidence module is considered the most important module in the Wang & Daniels proposed architecture because it plays an important role in analysis visualization of capture evidence.

Later, Wang & Daniels [10] proposed diffusion and graph spectral methods. These proposed methods aimed to establish a systematic forensics investigation process framework. Moreover, through these proposals Wang & Daniels attempted to provide high-performance computation methods to be used in the forensics analysis field as a form of well-utilized mathematical science.

In 2004, Gladyshev [11] proposed a formalized approach for Event Reconstruction. This approach was based on the terms of the finite state machine model of computation. The finite state machine model was used to define all possible attack scenarios in the computer network incidents. Furthermore, Gladyshev defined Event Reconstruction ’as a process of finding all potential computations of the machine that agree with the digital evidence of the incidentl;’. The scholar proposed an algorithm for the Event Reconstruction process that consists of three phases. The first phase calls for obtaining the finite state model of the computer system that is under the forensics investigation. In the second phase, all potential attack scenarios of the computer system incidents are defined by using the back trace method from the point in which the cybercriminal was discovered. The third phase calls for rejecting attack scenarios that conflict with the obtainable evidence [12].

Liu & Wijesekera [13] proposed merging sub-evidence graphs with an integrated evidence graph for network forensics analysis. This paper shows how to integrate different evidence graphs with or without the help of a corresponding attack graph. The proposal model assumes that an integrated evidence graph shows all attacks using global reasoning. Consequently, the research provided two algorithms that help integrate evidence graphs with a probabilistic evidence graph.

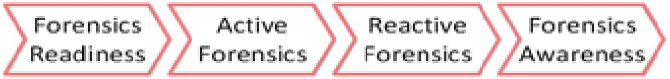


Fig. 1. Cybercrime Management Chain

Phillips & Swiler [14] proposed an approach for network risk analysis based on an attack graph that defines the set of attack paths that have a high probability of success for the attacker. This approach requires a predefined data-set as input information before starting to use the system. As a result, the system will generate an attack graph based on predefined information.

Sheyner et al. [15] proposed automated techniques in order to establish and generate the attack graphs. The techniques are based on a set of algorithms that are used to reconstruct attack scenarios automatically. After that, the reconstructed attack scenario is represented in the attack graphs. The visual representation of attack graphs allows forensics investigators to easily understand the attack scenario in an efficient manner. The authors implemented a network forensics tool based on the proposed algorithms, testing it in a small Local Area Network (LAN) that consists of an intrusion detection system and firewalls.

Bruaschi et al. [16] proposed a model that can organize digital forensics knowledge in a reusable way. In other words, this model can reuse the gathering techniques and some hypotheses in order to find the best guideline for hypotheses formulation.

III. Cyberspace Forensics Readiness and Security Awareness Model

In figure 2 shows the overall view of the proposed cyberspace forensic readiness and security awareness model.

Logs classification processes submodel: Basically, the op- erating system in the network firewalls and domain controllers (DCs) are able to classify the computer network and system events logs into predefined groups. This model was designed to increase the filtering process of the output events logs. It will classify the output logs into different groups, namely alerts and information.

Alert logs collection model: This model is designed to collect only the alert logs. These logs will be stored in the alert logs warehouse.

Alerts prepossessing model: The stored alert logs contain redundancy data and irrelevant information [11]. The alerts preprocessing model is used to filter out all redundancy data and irrelevant information from the alert logs. The alerts preprocessing model has two stages; format standardization and redundancy management. The format standardization process aims to convert the different event logs formats into one unified, common syntax format while the redundancy management process aims to reduce the duplication of the single event.

Assets Knowledge warehouse: Assets knowledge warehouse is designed to store basic information of all assets available in the network infrastructures.

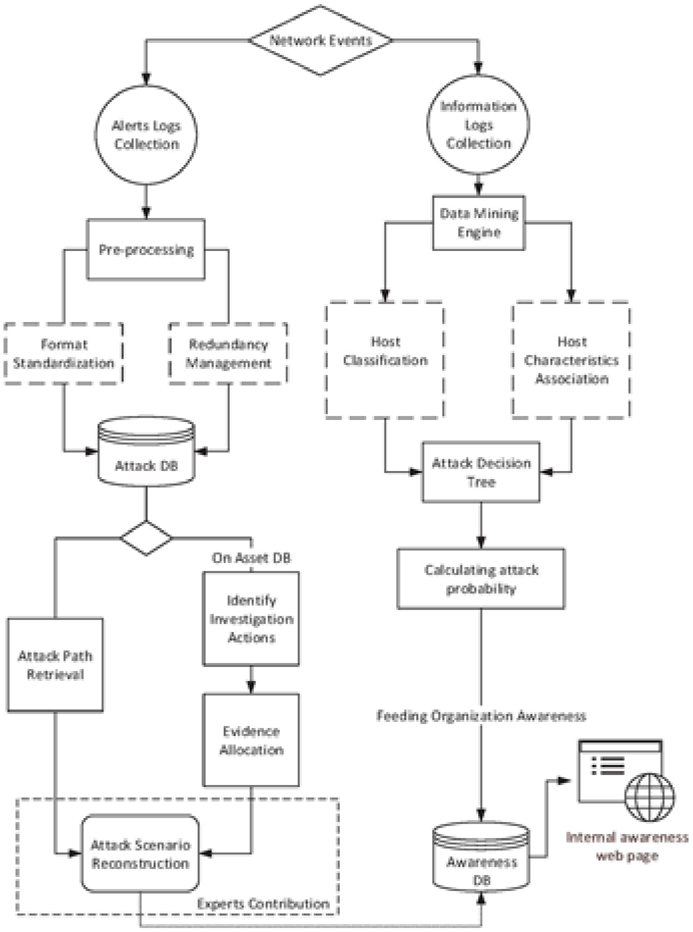


Fig. 2. Network Forensics Readiness and Security Awareness Framework

Attack knowledge warehouse: The assets knowledge warehouse is designed to store basic information of all assets available in the network infrastructures.

Attack path Retrieval: An attack graph provides a visual representation of the attack paths as well as evidence for each node (host) in each path (see figure 4). The attack paths describe all exploited network assets. The attack graphs will be generated based on databases, namely asset knowledge and attack knowledge. The nodes indicate the exploited hosts while the edges indicate the security vulnerability used to hack the host. The information shown in this graph is based on a chain of custody manner.

Scenario reconstruction submodel: After generating the attack and evidence graphs, this model is used to reconstruct the attack scenario. This process will reprocess the criminal graph with the help of criminology sciences and hypothesis expert knowledge.

A. Information prepossessing model

Information collection sub model: The information collection submodel will collect all output of information logs from event log classification processes and forward it to the information logs warehouse.

TABLE I. THE EVENTS STRUCTURE

The Events structure field	
Field	Description
Type	Shows the type of events (Information, Warning, Error, etc.)
Time	Shows the time of the event happened
Date	Shows the date of the event happened
Event ID	Shows an event log number that identifies the event type
Device	Shows the device where the event happened
User	Shows the computer user who has generate the events or who logged to the computer system when the happened
Source	Shows the source produced the event

Data mining Engine: As there are so many information logs in the information logs warehouse, it is very difficult to check all of them and update information security awareness. This step is used to convert information logs into an easier format that will be useful for security information awareness.

The data mining engine consists of two types of processes; host classification types and host characteristics associations. First, the host classification process will be used to classify all existing assets in the network infrastructure into certain groups based on host types, such as router, switches, domains controllers, firewalls, etc. Second, host characteristics associations will be used to associate each log to the appropriate predefined group. Using the association process, it will be necessary to analyze the logs header format to be able to know the appropriate predefined group.

Calculating attack probability: Calculating the attack probability process will be used to process the output results of the attack decision tree. This process examines the attack probability for each asset (for example, the file server probability affected by DOS) based on previous experiments knowledge through the data-set analysis.

Awareness DB: The awareness DB is used to store the attack probability for each asset in the network infrastructure. This database feeds the internal awareness Web page through security awareness and vulnerabilities for network assets.

B. The normalization process of alerts and information logs

As mentioned earlier, the normalization process will be used to convert the event logs formats into a unified format. This process will help to aggregate the logs and reduce redundancy and noise information. Table I I shows a proposed unified structure field of event logs.

C. The relationship between the evidence and attack

It is very important to know the relationship between the evidence and attacks. This relationship helps us in the investigation process, as well as increases the admissibility of the investigation case in court. Moreover, the increased amount of evidence related to a specific attack case will increase the background information about the attacker. There are different types of relationships between the detected evidence and the attacks, including one-to-one, one-to-many, and many-to-one.

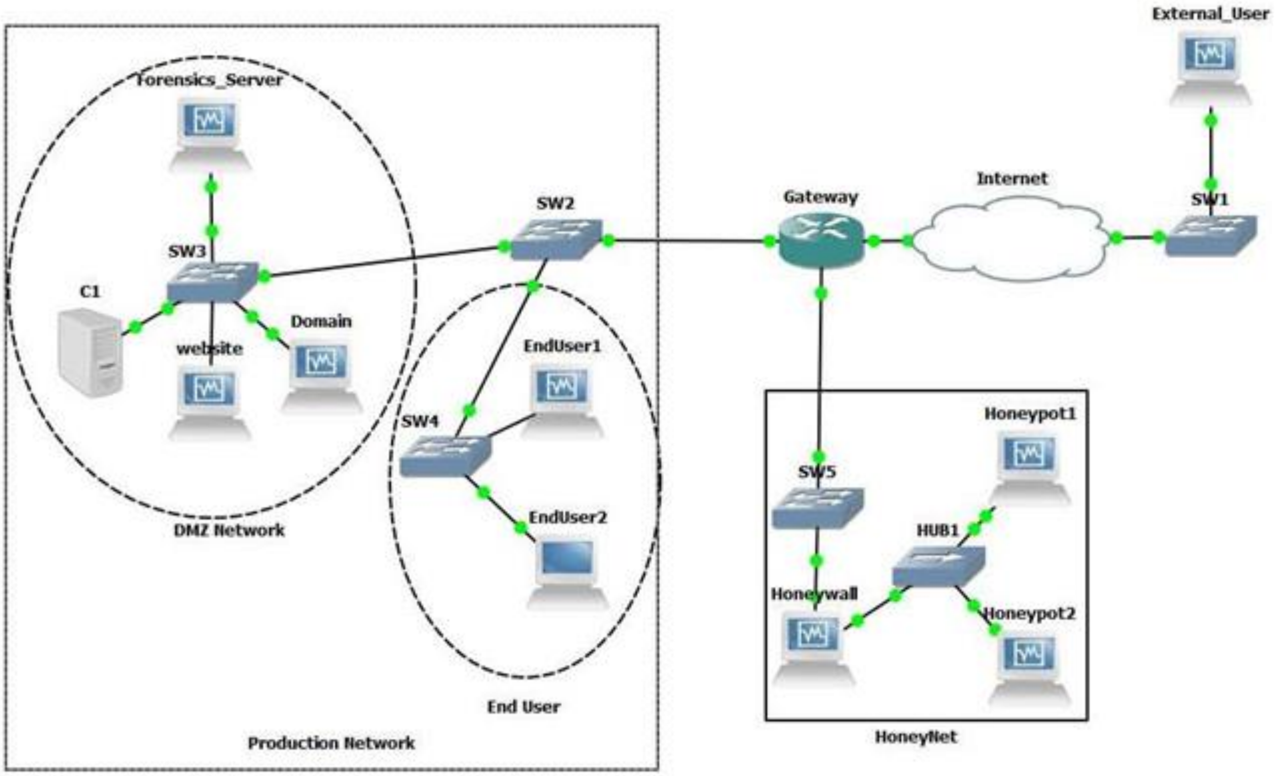


Fig. 3. Simulation Honeynet Network in GNS3

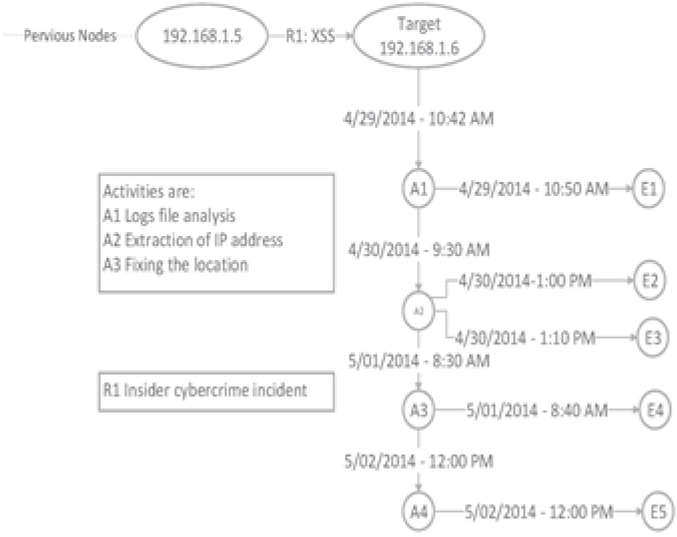


Fig. 4. Integration between Attack Graph and Evidence Graph

IV. Example of Criminal Graph

This paper presents a picture of the proposed graph that integrates the attack and evidence graphs. Section IV-A establishes a criminal scenario.

A. Simulating SQL-Injection Cyber-attack using GNS3

The authors [17] presented a simulation study network attack scenario. This is the first step towards validating the proposed model. The simulation case study used capturing, normalizing and analyzing events that are introduced in section III. The main point of designing virtual network attack environments is to create a sandbox that allows one to perform such experiments, from real assets and at a low cost. Both the capturing and examination of the events were conducted in the simulated case study. The detection of network artifices changes after the execution of SQL-Injection attacks were also recorded. The outcome of this experiment can be used as a recommendation in real cyber-infrastructure. The core idea of the case study is to examine the website that has been compromised by an SQL injection attack. To simulate this attack scenario many open source tools were used such as Graphical Network Simulator (GNS3), Oracle VM Virtual Box and VMWare workstation. The wireshark forensics tool was also used to detect criminal activity from the network layer (Layer 3 in OSI model), in addition, the victims and attackers devices by using the Volatility Framework 2.4 were also examined.

Simulation approaches helps to graphically simulate an attack for courts, Jury and Investigators. The simulation approaches also helps to simplify the incidence (1 Image = 1000 words). The current study [17] proposes Investigation learning methodology based on the proposed case study. The learning methodology consists of two stages; stage one is to build a network topology of the proposed case study and stage two is to create a network union Matrix.

This approach allows specific network devices configuration to be simulated, perform SQL injection attacks against victim machines and collect network logs. The main motivation of thiswork is to finally define an attack pathway prediction methodology that makes it possible to examine the network artifacts collected in case network attacks.

Figure 4 shows the integration between the attack graph and the evidence graph. Moreover, nodes indicate compromised hosts while edges refer to security vulnerabilities used by the attacker. Moreover, under each compromised host, there is another graph that shows the series of actions carried out by the forensics investigators. Furthermore, the actions carried out by the forensics investigation are linked to another graph called the evidence graph. This graph will show the output evidence as a result of each forensics investigative action.

The authors purposed a new network forensics model [8] that can makes network events admissible in the court of law. The present model collects available logs from connected network devices, applies decision tree algorithm in order to filter anomaly intrusion, then re-route the logs to a central repository where events management functions are applied.

V. Conclusions and Future Work

This proposed model contains approximately fifteen different models. The proposed models work as a single unit in order to process and normalize the captured network logs. The main point of designing the model is to find a way to forensically visualize the evidence and attack scenario in a computer system. Moreover, this paper listed some methods and approaches proposed by scholars to construct the attack scenario. Nevertheless, the graph representation is one of the best approaches used in the forensics investigation; the researchers in this field have proposed several types of graphs, including scenario graphs, logic exploitation graphs, forensics graphs, attack graphs, and evidence graphs.

Since the attempt to reconstruct scenarios of network attacks from collected data (i.e., alarms, alerts and logs) requires brain-like reasoning to understand these events. Therefore, Bio-inspired approaches [18] to self-organizing network events and creating the linkage between them are of relevance to the research studies. The future plans is to examine the possibility to replace the traditional database approach to storing events with a bio-inspired mechanism and, study the affect of that on the quality of the scenarios produced.

This model acts as a first step toward network logs analysis. The future work will focus on involving mathematics and algorithm science for each proposed blocks to help validate the model. Furthermore, trying to utilize criminology science to enhance any future proposed models or approaches are a key priority.

Acknowledgment

The authors would like to acknowledge the support of by the Insight Centre for Data Analytics and UCD Centre for Cybersecurity and Cybercrime Investigation.

References

[1] I. Baggili and M. Kiley, “Digital forensics a brief overview of critical issues,” Digital Forensics Investigator News, 2008.

[2] G. Palmer, “A road map for digital forensics research-report from the first digital forensics research workshop (dfrws),” Utica, New York, 2001.

[3] A. Salim Al-mahrouqi, S. Abdalla, and T. Kechadi, “E-government alerts correlation model,” in Qatar Foundation Annual Research Conference, no. 1, 2014, p. ITPP1120.

[4] V. Leucari. (2005) Analysis of complex patterns of evidence in legal cases: Wigmore charts vs. bayesian networks. [Online]. Available: https://www.ucl.ac.uk/jdi/research/evidence-network/docs/BURGLARY.PDF

[5] L. P. Swiler, C. Phillips, D. Ellis, and S. Chakerian, “Computer-attack graph generation tool,” in DARPA Information Survivability Conference &amp; Exposition II, 2001. DISCEX’01. Proceedings, vol. 2. IEEE, 2001, pp. 307–321.

[6] S. Neralla, D. L. Bhaskari, and P. Avadhani, “A novel graph model for e-mail forensics: Evidence activity analysis graph,” International Journal of Engineering Science and Technology, vol. 5, no. 10, p. 1750, 2013.

[7] J. James, P. Gladyshev, M. T. Abdullah, and Y. Zhu, “Analysis of evidence using formal event reconstruction,” in Digital Forensics and Cyber Crime. Springer, 2010, pp. 85–98.

[8] A. Al-Mahrouqi, S. Abdalla, and T. Kechadi, “Efficiency of network event logs as admissible digital evidence,” in Science and Information Conference 2015, London, United Kingdom, 28-30 July 2015, 2015.

[9] W. Wang and T. E. Daniels, “Building evidence graphs for network forensics analysis,” in Computer Security Applications Conference, 21st Annual. IEEE, 2005, pp. 11–pp.

[10] ——, “Diffusion and graph spectral methods for network forensic analysis,” in Proceedings of the 2006 workshop on New security paradigms. ACM, 2006, pp. 99–106.

[11] P. Gladyshev, “Formalising event reconstruction in digital investigations,” Ph.D. dissertation, University College Dublin, 2004.

[12] M. Sebastian and P. Chandran, “Towards designing a tool for event reconstruction using gladyshev approach,” in Proceedings of the 2011 ACM Symposium on Applied Computing. ACM, 2011, pp. 193–194.

[13] A. C. Liu and D. Wijesekera, “Merging sub evidence graphs to an integrated evidence graph for network forensics analysis,” Advances in Digital Forensics IX, pp. 227–241, 2013.

[14] C. Phillips and L. P. Swiler, “A graph-based system for network-vulnerability analysis,” in Proceedings of the 1998 workshop on New security paradigms. ACM, 1998, pp. 71–79.

[15] O. Sheyner, J. Haines, S. Jha, R. Lippmann, and J. M. Wing, “Automated generation and analysis of attack graphs,” in Security and privacy, 2002. Proceedings. 2002 IEEE Symposium on. IEEE, 2002, pp. 273–284.

[16] D. Bruschi, M. Monga, and L. Martignoni, “How to reuse knowledge about forensic investigations,” in Digital Forensics Research Workshop, 2004, pp. 10–13.

[17] A. Al-Mahrouqi, P. Tobin, S. Abdalla, and T. Kechadi, “Simulating sql-injection cyber-attacks using gns3,” International Journal of Computer Theory and Engineeringl, vol. 8, no. 3, pp. 213–217, 2015.

[18] D. Floreano and C. Mattiussi, Bio-inspired artificial intelligence: theories, methods, and technologies. MIT press, 2008.

Multi-Biometric Systems: A State of the Art Survey and Research Directions

Ramadan Gad
Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University,
Egypt.

AYMAN EL-SAYED
Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University,
Egypt.

Nawal El-Fishawy
Computer Science and Engineering,
Faculty of Electronic Engineering, Menoufia University,
Egypt.

M. Zorkany
Electronic and Communication Engineering,
National Telecommunication Institute
Egypt.

***Abstract*—Multi-biometrics is an exciting and interesting research topic. It is used to recognizing individuals for security purposes; to increase security levels. The recent research trends toward next biometrics generation in real-time applications. Also, integration of biometrics solves some of unimodal system limitations. However, design and evaluation of such systems raises many issues and trade-offs. A state of the art survey of multi-biometrics benefits, limitations, integration strategies, and fusion levels are discussed in this paper. Finally, upon reviewing multi-biometrics approaches and techniques; some open points are suggested to be considered as a future research point of interest.**

Keywords—Biometrics; Multimodal biometric systems; fusion levels; recognition methods; authentication

I. INTRODUCTION

Authentication (identifying an individual using security system) of users is an essential but, difficult accurate and secured practical authentication technology. Traditional techniques for user authentication could be categorized as [1, 2]: (1) Token based techniques (i.e. key cards and smart cards) and (2) Knowledge-based techniques include text-based and picture-based passwords (often mix of username and password).

Due to vulnerabilities in above methods (It could be easily transgressed or lost or forgotten); Traditional techniques are considered to be not reliable or secure, and are not presently sufficient in some security application zones [3, 4]. The primary advantage of biometrics over these methods is that it cannot be misplaced, forgotten or stolen. Also, it is very difficult to spoof biometric traits . Due to greater accuracy and higher robustness of biometric recognition [1, 5]; Biometric solutions become popular and preferred methods to analyze human characteristics for security - authentication and identification - purposes[6]. It could not be duplicated or counterfeited and misused.

Practically, the use of biometrics information is the most secure method [7]. Consequently, it is now needed in many fields such as surveillance systems, security systems, physical buildings [8]. Other applications of biometrics systems include [9, 10]: access control (access to computer networks), forensic investigations, verification and authentication, e-commerce, online banking, border control, parenthood determination, medical records management, welfare disbursement and security monitoring. Biometrics applications increased dramatically in functionality in many more fields.

In the most general definition, "Biometric technologies" is defined as an automated methods of verifying and/or recognizing the identity of a living individual based on two categories : (1) *Physiological biometrics* include (Facial, hand and hand vein infrared thermogram, Odor, Ear, Hand and finger geometry, Fingerprint, Face, Retina, Iris, Palm print, Voice, and DNA) [10], and (2) *Behavioral biometrics* like (Gait, Keystroke, Signature) which measure the human actions [8]. Also, human electrocardiogram (ECG) signal is considered one of Biometric features used in individual recognition and authentication[11].

Depending on the application context, biometric systems may operate in two modes: verification mode and identification mode [5]. Through *verification mode*, the system verifies the identity by comparing the enrolled biometric trait by a stored biometric template in the system (1:1). This mode is used for positive recognition, and it aims to prevent the multiple individuals from using the same identity. In the *identification mode*, the enrolled sample is then compared with existing templates in a – central – database (1: M) . A database search is crucial and needed. The identification mode is critical in negative recognition applications, which aims to prevent a single user from using multiple identities [12]. Negative identification is also known as screening [8]. Obviously, verification is less computationally expensive and more robust compared with identification. On the other hand, the latter is more convenient and less obtrusive [13].

Multi-biometric systems distinguished over traditional uni-biometric systems as it [14] addresses the issue of non-universality and noisy data. Multi-biometric systems can facilitate the indexing of large-scale biometric databases. Also, it becomes not easy for an impostor to spoof all the biometric traits of an authorized enrolled person. Generally, It is much more vital to fraudulent technologies because it is more difficult to forge multiple biometric characteristics. Multi-biometric recognition systems also have benefits in the continuous monitoring of an individual in situations or tracking him when a single trait is not sufficient in use. These systems continue to operate even if part of biometric sources become unavailable of a failed (i.e. sensor malfunction, software malfunction, or deliberate user manipulation); it may view as a fault tolerant system. For these benefits, multimodal expected to provide higher accuracy rate.

The rest of this paper is organized sequentially as follow: Section II will overview the biometrics characteristics followed by section III to discuss the unimodal biometrics' drawbacks. Next, Section IV will discuss the multi-biometrics advantages and limitations, categories, and integration scenarios. After that, section V is to discuss biometrics quality performance and metrics. different fusion levels before and after matching, depended on theses metrics, will be discussed in section VI. Benefits and drawbacks for each approach will be declared with evidence of previous research. Moreover, section VII will show the design issues and trade-offs related to any multi-biometric recognition system. Finally, Section VIII suggests some open points for further investigation and research.

II. Biometrics Overview

A biometric system to be practical and reliable should meet the specified requirements/characteristics [15] [4]: *Universality (availability)*, each person should have the characteristic. Availability is measured by the "failure to enroll" rate. *Distinctiveness:* It declares that any two persons should sufficiently have different characteristic. It is measured by the False Match Rate (FMR), also known as "Type (II) error". *Permanence (robustness)*, the characteristic should be stable (with respect to the matching features) over a period of time. Which means the stability over age. Robustness is measured by the False Non-Match Rate (FNMR), also known as "Type (I) error" . *Collectability (accessible)*, the characteristic can be measured quantitatively, and easy to image using electronic sensors. Accessibility can be quantified by the "throughput rate" of the system. *Performance:* It means to achieve recognition accuracy, speed, and the resources required to the application. *Acceptability,* The particular user population and the public, in general, should have no (strong) objections to the measuring/collection of the biometric characteristic. Acceptability is measured by polling the device users . *Resistance to Circumvention*, tests and proofs how the system resists fraudulent methods easily.

Consequently, a brief comparison of the most known biometric techniques based on above factors are shown in table (I) [12, 16], to differentiate between the biometrics modalities as a unimodal trait.

Which biometric characteristic is best? Each biometric feature has its own strengths and weaknesses and the choice typically depends on the application. Accordingly, each one could be used in authentication and/or identification applications [17]. Predicting the "false acceptance" and "false rejection" rates, system throughput, user acceptance, and cost savings for operational systems from test data, is a surprisingly difficult task.

Consequently, it is impossible to state that a single biometric characteristic is "best" for all applications, populations, technologies and administration policies.

TABLE I. Comparison of Biometric Characteristics [12, 16]

Biometric Characteristic	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
Facial Thermogram	H	H	L	H	M	H	L
Hand Vein	M	M	M	M	M	M	L
Gait	M	L	L	H	L	H	M
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Ear	M	M	H	M	M	H	M
Hand Geometry	M	M	M	H	M	M	M
Fingerprint	M	H	H	M	H	M	M
Face	H	L	M	H	L	H	H
Retina	H	H	M	L	H	L	L
Iris	H	H	H	M	H	L	L
Palm Print	M	H	H	M	H	M	M
Voice	M	L	L	M	L	H	H
Signature	L	L	L	H	L	H	H
DNA	H	H	H	L	H	L	L

[a.] (H: High, M: Medium, and L: Low)

III. Unimodal Biometrics Limitations

Any single modal biometric has limitations. For example, iris recognition suffers from some problems like camera distance, eyelids and eyelashes occlusion, lenses, and reflections [18-20]. Face changes overages and unstable, and twins may have similar face features. Also, fake faces from mobiles as example, and masks used to attack the system . Fingerprint may have some cuts, burns, and small injuries temporary or permanent . Moreover, fake fingers made from gelatin and/or silicon have ability to attack the fingerprint-based recognition system . Cold leads to voice problems and a tape recordings may be used to hack the system [13]. The fingerprint of DNA needs several hours to be obtained. Besides, DNA includes sensitive information related to genetic of individuals and the test is quite expensive to perform . Hand geometry is not distinctive enough to be applied to a large population. Thus, it is not suitable for purpose of identification [16]. Gait is sensitive to body weight and not stable; it is not used for large population and not reliable enough . Signature is not universal and changes with time. Offline ones are forgery while, Online signature cannot applied for documents verification (i.e. Government documents and bank cheques) . None of above traits alone can ensure perfect recognition performance. Nevertheless, the biometric system (either an 'identification' system or a 'verification' system) can also be attacked by the outsider or unauthorized person at various points [21]. Combining multiple modalities is a good idea to decrease these conditions.

The unimodal biometric rely on the evident single source of information for authentication (e.g., single fingerprint, face) . Single modal biometric traits may not achieve the desired performance requirements; as they have plenty of error rates [5, 15]. These systems have to contend with a variety of problems such as:

- *Noise in sensed data;* defective or improperly maintained sensors (i.e. accumulation of dirt on a fingerprint sensor) could produce deformed and noisy data. For instance, a cold has effects on the voice, wearing glasses alters iris recognition performance, variations in light or illumination in face sensed …etc.

- Distinctiveness (Intra-class variations and Inter-class similarities); Biometric trait is expected to be varied significantly across two persons. Intra-class variations occur when a user interacts with the sensor incorrectly (e.g., incorrect facial pose). Also, characteristics of the individuals are formed with the large inter-class similarity (overlap) in the feature sets of multiple users.

- Non-universality; means the non-ability of the biometric to acquire meaningful biometric data from a group of users due to the poor quality and consistency of the acquired biometric data as a result to error or a fault in the sensor. For example, many of population (about 4%) may have scars or cuts in fingerprints. As a result, a fingerprint biometric system, may extract incorrect minutiae features from them. Also, user-sensor interaction is adjustment incorrectly. Of course, this may give undesired matching result.

- Spoof attacks; a fake traits or biometrics of the authorized user are enrolled and saved in the template database; an imposter person may attempt to spoof these sensed data when the traits are used. As in [22], artificial fingers/fingerprint can be used to spoof the verification system. This type of attack is common when using behavioral characteristics.

On behave of above problems, unimodal biometric systems suffer other drawbacks like: insufficient population coverage, lack of individuality, lack of invariant representation, and susceptibility to circumvention [7].

These problems lead to higher False Reject Rate (FRR) and False Accept Rate (FAR) [4, 10, 23] as will be shown later in quality metrics, in section 5.

IV. Multi-Biometrics as a Solution

Biometric fusion has a history of more than 30 years . More than one biometric combined to investigate high performance multi-biometric recognition system. Multi-biometrics has addressed some issues related to unimodal this make it has some benefits over unimodal biometrics such as recognition accuracy, privacy, and biometric data enrollment.

Recognition accuracy: Its accuracy is better as compared to the unimodal biometric system [24]. The multi-biometric system is expected to be more accuracy and reliability due to the multiple, biometric traits independency, and difficult to forge all of them [5, 10]. As the combination of each of the biometric identifiers offers some additional evidence about the authenticity of an identity claim, one can have more confidence in the result. For example, two persons may have the similar signature patterns, in which case, the signature verification system will produce large FAR for that system. Addition of face recognition system with the signature verification system may solve the problem and reduce the FAR [9]. Experiments have shown that the accuracy of multimodality can reach near 100% in identification.

Privacy: Multimodal biometric systems increase resistance to certain type of vulnerabilities. It prevents from stolen the templates of biometric system as at the time it stores the two characteristics of biometric system in the database [25]. For example, it would be more challenge for attacker to spoof many different biometric identifiers[9]. Further, when two or more modalities are used for authentication, it leads to become not easy to spoof the biometric system.

Biometric data enrollment: Multimodal biometric systems can address the problem of non-universality. In case of unavailability or poor quality of a particular biometric data, other biometric identifier of the multimodal biometric system can be used to capture data. For example, a face biometric identifier can be used in a multimodal system (involves fingerprint of general labors with lots of scars in the hand) [9]. It makes better system operation [24]. Multi-biometric system also addresses the problem of noisy data effectively (i.e. illness affecting voice, scar affecting fingerprint). They allow indexing or filtering of large biometric databases, and are robust to noise. Thus, it provides universal coverage and improves matching accuracy [10, 15, 26].

A. *Multimodal Categories*

Multi-biometric systems have two basic categories: synchronous and asynchronous. In synchronous, two or more biometrics combined within a single authorization process. On the other hand, asynchronous system uses two biometric technologies in sequence (one after the other) [27]. Multimodal biometric systems can operate in three different modes [5]:

- *Serial Mode (cascade mode)* – each modality is examined before the next modality is investigated. The overall recognition duration can be decreased, as the total number of possible identities - before using the next modality - could be reduced

- *Parallel Mode* – sensed/captured data from multiple modalities are used in concurrent way to perform recognition. Then the results are combined to make final decision.

- *Hierarchical Mode* – individual classifiers are combined in a hierarchy -tree like- structure. This mode is preferred when a large number of classifiers are expected.

B. *Multi-Biometrics Integration Scenarios*

Recognition systems using multiple biometric traits are designed to operate in one of the integration scenarios as below:

1) *Multi-sensor systems*

The information of the same biometric obtained from different sensors are combined for all. For example, complementary information corresponding to fingerprints can be acquired using different types of sensors (like optical and capacitive sensors). Information obtained are then integrated using sensor level fusion technique[15].

2) Multi-modal systems

More than one biometric trait is used for user identification. For example, the information obtained using face and voice features or other can be integrated to establish the identity of the user[27]. This can be more costly; because it requires multiple sensors with each sensor sensing different biometric characteristics. But, the improvement in performance is substantial.

3) Multi-instance systems

Multiple instances of a single biometric trait are captured. For example, images of the left and right irises can be used for iris recognition. Also, fingerprints from two or more fingers of a person may be combined or one image each of the same person may be combined. If a single senor is used to acquire these images in a sequential manner, the system can be made really cost effective, as it does not require multiple sensors. Moreover, it does not incorporate additional feature extraction and matching modules [17].

4) Multi-sample systems

Multiple samples of a same biometric trait are used for the enrollment and recognition. For example, along with the frontal face, the left and right profiles are also captured. Multiple impression of the same finger, and multiple samples of a voice can be combined. Multiple samples may overcome poor performance. But, it requires multiple copies of sensors, or the user may wait a longer period of time to be sensed or a combination of both[15].

5) Multi-algorithm systems

Multiple different approaches to feature extraction and matching algorithms are applied to a single biometric trait. Final decision obtained if any of the matching fusion technique can be applied on the results obtained using different matching algorithms. These systems are more economical as no extra device is required to capture the data. But, these are more complex because of application of different algorithms[15].

6) Hybrid systems

It is a system which integrates more than one of the above mentioned multi-biometric systems. For example, two face recognition algorithms can be combined with two fingerprint recognition algorithms. Such a system will be multi-modal and multi-algorithmic system. Moreover, if multiple sensors are used to obtain these images, then it will be multi-sensory, and if multiple instance of the finger is used, it will be multi-instance system also.

Both of hybrid systems and multi-modal systems can be desired by using multiple modalities. However, the rest can be achieved with the only help of even single modality [23]. The different types of multi-biometric are shown in figure (1).

C. Limitation of Multi-biometrics System

Some lacks are still found such as noise in the biometrics like scratches in the fingerprint and lens mark in iris, this will lead to increase the (FRR). Moreover, the accuracy of the multi-biometric enrollment and multi-biometric identification need to be improved. In multi-biometrics, failure of one biometrics will make the whole system to fail [28]. In addition, multimodal biometric systems, may be more expensive and complicated due to the requirement of additional hardware and matching algorithms, and there is a greater demand for computational poser and storage [9]. Recent research has revealed that multi-biometric systems can increase the security level as a means to enhance network security to people who are encouraged to use biometric systems in this field. However, it need more efforts and research to face some types of attacks such as: spoof attack, replay attack, substitution attack, Trojan horse attack, transmission attack, template database attack, and decision attack [17]. Next section will list the performance metrics that distinguish between the multi-biometrics techniques.

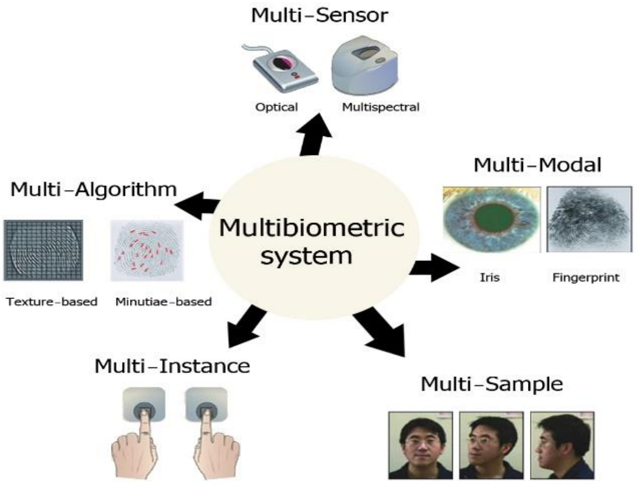


Fig. 1. The different types of multi-biometric system. [15]

V. QUALITY PERFORMANCE AND METRICS

Various quality performance metrics measure the performance of any biometric authentication techniques. It helps comparing systems and motivating the progress [13]. The most common performance metrics of biometric systems are described below [12] :

False Accept Rate (FAR) or (False Match Rate (FMR)): Mistaking the biometric measurements from two different persons to appear as if they are from the same person due to large inter-user similarity. It measures the percent of invalid matches. The FAR is defined as in (1) [1, 29, 30]:

$$FAR\% = \frac{T_{Faccept}}{T_{Fsubmit}} \times 100 \qquad (1)$$

Where, $T_{Faccept}$ is total number of forgeries accepted and $T_{Fsubmit}$ is total number of forgeries submitted to the system test. In a good authentication system this rate must be low.

False Reject Rate (FRR) or (False Non-Match Rate (FNMR)): Mistaking two biometric measurements from the same person to appear that they are from two different persons due to large intra-class variations. It measures the percent of valid inputs being rejected. The FRR is defined as in (2) [24]:

$$FRR\% = \frac{T_{Greject}}{T_{Gsubmit}} \times 100 \qquad (2)$$

Where $T_{Greject}$ is the total number of genuine test pattern rejected, and $T_{Gsubmit}$ is total number of genuine test submitted to the system. This must be low to achieve good Performance. The average of the FRR and FAR is called the Average Error Rate (AER)[29].Genuine Acceptance Rate (GAR) sometimes used, which is the percentage of the likelihood that a genuine individual is recognized as a match [8]. GAR of a valid user can be obtained by equation (3) [31].

$$GAR\% = 1 - FRR\% \qquad (3)$$

Equal Error Rate (EER): For a simple empirical measure, it is used to summarize the performance of a biometric system that is defined at the point where False Reject Rate (FRR) and False Accept Rate (FAR) are equal . System with the lower EER, is the more accurate and precise [1, 9, 30]. The EER is also called the type (III) error [29].

Failure to Capture (FTC): denotes the percentage of times the biometric device fails to automatically capture a biometric characteristic when presented correctly. This usually happens when system deals with a signal of insufficient quality [24].

Failure to Enroll Rate (FER or FTE): denotes the percentage of times users cannot enroll in the recognition system[32]. Data input is considered invalid due to poor quality.

Template Capacity: It is the maximum number of sets of data which can be input in to the system [24].

Usually, the above performance metrics are expressed using different graphs such as Receiver Operating Characteristic (ROC), Score Histogram (SH), and Cumulative Match Characteristic (CMC) [9]. *Receiver Operating Characteristic (ROC) curve*: There is a trade-off between FAR and FRR in every biometric system. In fact, both of them are functions of the system threshold (t); if it is declined to make the system achieves higher tolerance to input variations and noise, then FAR increases. On the other hand, if it is raised to make the system more secure, then FRR increases accordingly . The ROC plot is obtained by graphing the values of FAR against FRR, at various operating points (thresholds) on a linear or logarithmic or semi-logarithmic curve. Detection Error Trade off (DET) is a common variation, which is obtained via normal deviate scales on both axes [24]. This graph is more linear that illuminates the differences for higher performances. *Cumulative Match Characteristic (CMC) curve*: is used in biometric identification to summarize the identification rate at different rank values [8]. *Score Histogram (SH)*: plots the frequency of the scores for matches and non-matches over the match score range. These metrics are needed to differentiate between each level fusion and method considered for the multi-biometrics as a solution. Categorization of different levels of fusion will be discussed in next section.

VI. Levels of Fusion in Multimodal Biometrics

Multimodal biometric fusion combines the distinguished aspect from different biometric features to support the advantages and reduce the drawbacks of the individual aspects [5]. The fundamental issue of information fusion is to determine the type of information that should be fused and the selection of method for fusion . The goal of fusion is to devise an appropriate function that can optimally combines the information rendered by the biometric subsystems [8].

In multimodal biometrics, the fusion scheme can be classified as sensor level, feature level, match score level, rank level, and decision level [4] as shown in figure (2). The process can be subdivided into two main categories: prior-to-matching fusion and after matching fusion [33]. Figure (3) [9], shows these fusion levels possibilities at each module. The hybrid one is mixing two or more from these level fusions.

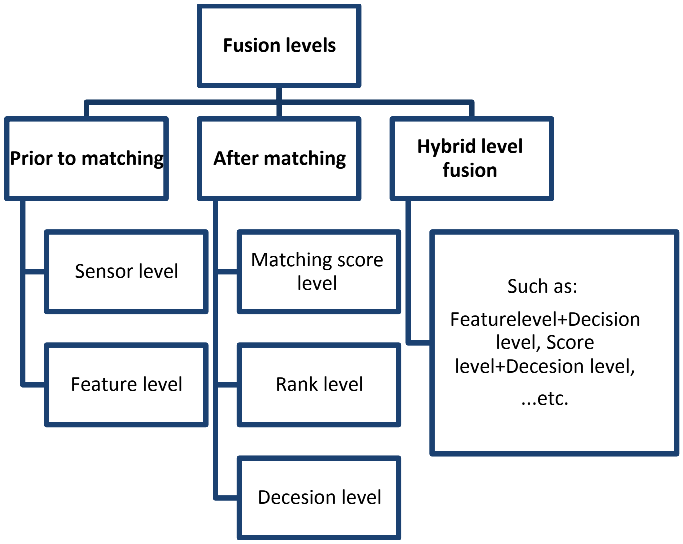


Fig. 2. Categories of different fusion levels

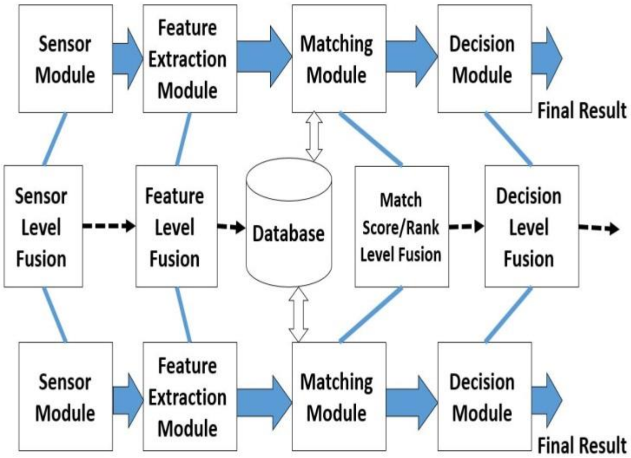


Fig. 3. Prior-to-matching and after matching fusion levels related to biometric system modules [9]

A. Prior to Matching Fusion

Fusion in this category integrates evidences before matching. This can be classified into two different categories as follows:

1) Sensor level fusion

Principles- A new biometric data generated by merging the raw data obtained from multiple sources. Then, trait can be extracted. A single sensor or different compatible sensors like fingerprint, iris scanner, etc., represents the samples of the single biometric trait sensed [23]. This level of fusion is also known as data level fusion or image level fusion (for image based biometrics) [4].

Discussion- Sensor level fusion can benefit multi-sample systems which capture multiple snapshots of the same biometric [15]. Compared to other fusion types, it has a lot of information. It is projected to improve the recognition accuracy. Sensor fusion addresses the problem of noise in sensed data because improper maintenance of sensors [4]. However, raw images are either not available or the information available from the different sources is not compatible. For this unavailability and incompatibility of desired information, sensor level and feature level fusion are not possible in all cases [9]; Very less work has been done in this type of fusion [17]. As an example of sensor level fusion, Ratha et al. [34] described a fingerprint mosaicing scheme to integrate multiple snapshots of a fingerprint as the user rolls the finger on the surface of the sensor.

2) *Feature level fusion*

Principles- The correlated feature sets extracted from different biometric channels (modalities) can be fused by using specific fusion algorithm forming a composite feature set, passed to the matching module [5, 27]. This done after normalization, transformation and reduction schemes [33]. The goal of feature normalization is to modify the location (mean) and the scale (variance) of the feature value via a transform function in order to map them into a common domain. (e. g. Min-max normalization, Median normalization...etc.) . Transformation or Feature Selection is algorithm use to reduce the dimensionality of the feature set. (e. g. Sequential forward selection, Sequential backward selection, Principal Component Analysis (PCA), etc.) [15].

Discussion- Final feature vectors could be either homogeneous or heterogeneous. The feature sets are from different algorithm and modalities; so the consolidation of feature set may have some problems [5, 23]. The relationship between these features of different biometric systems may not be well known, and structurally incompatible features are common. In addition, concatenating two feature vectors might lead to the dimensionality problem [4]. Lead to these difficulties, fusion level reported in limited research work.

For example, in year 2004, Feng et al. [19] developed a feature level fusion based multimodal biometric system using face and palm print. They used Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as classification algorithms. The PCA-based accuracy rate was (70.83%, 85.83%) for (face, palm print), while 95.83% after fusion. Moreover, ICA-based accuracy rate was (85%, 92.5%) for (face, palm print), while 99.17% after fusion. some previous fused modalities based on feature level fusion as in [35-41].

B. *After Matching Fusion*

Prior to matching fusions sometime don’t involve multiple modalities. Also, the fusion of data set is more complex, and it is not good to ignore any data [23]. After matching fusion integrates evidences of after matching module. This can be classified into three different categories:

1) *Matching score level fusion*

Principles- Individually, Extracted feature vectors (generated separately for each modality) are compared with the templates enrolled in the database for each biometric trait in order to generate the match scores [5]. Output set of match scores are fused to create composite matching score (single scalar score) [4]. This fusion technique is also known as confidence level or measurement level fusion. Density, transformation, and classifier based score fusion are different methods to achieve this fusion level [23].

The matching scores cannot be used or combined directly; because these scores are from different modalities and based on different scaling methods. Score normalization are required, by converting the scores into common similar domain or scale. This can be carried out with different methods. Slobodan Ribaric and Ivan Fratric discovered - piecewise linear normalization - new normalization technique. Their experiments used palm print and facial features.

Discussion- Applying fusion at this level is preferred as it is easy to obtain and combine matching scores of different biometrics [10]. It provides richest set of information about the biometric data. But complexity is more [23]. A lot of work has been done using match score level fusion. It is the most investigated fusion method so far which considers the match or similarity/distance score for fusion. But, the similarity/distance scores need to be normalized before fusion (as they can be in different ranges) [9]. Choosing inappropriate normalization technique result leads to very low recognition performance rate [4].

As an example, face modality and hand modality match scores together are combined in paper. Also, the match scores generated by the face, fingerprint and hand modalities of a user combined via the simple sum rule to obtain a new match score, after that it is used to make the final decision [18]. A rest of some previous work in [15, 18, 26, 33, 34, 42-57], considers matching score level fusion.

2) *Rank level fusion*

Principles- In this new fusion approach, each classifier associates a rank with each enrolled trait to the system (a higher rank indicating a good match). It consolidates multiple unimodal biometric matcher outputs, and determining a new rank that would help in estimating the final decision [4, 5]. Generally, the rank level fusion is adopted for the identification rather than verification. Here, the working procedures are: first, generate a rank of identities sorted with all modalities. Second, by help of any method of fusion, the ranking for each individual available for different modalities fused. Finally, the identity with the lowest score is the correct identified one [23].

Discussion- Beside it orders the identities based on those similarity/distance, it does not need any normalization procedure [9]. This method provide more accuracy comparing with just identifying best match with one modality. Unlike match score level fusion, it is easily possible to compare the ranking from different biometric modalities. As a result, it is so easy to make the decision [23].

However, this type of fusion has one weakness. In a case of multimodal biometric, which more different identities output from number of matching modules appear some identities of only one matcher, a wrong results act a risk of achieving the rank level fusion [33]. Unlike to match score level fusion, rank level fusion provides less information. It is better, because it provides a rank to different matches and also weights can be assigned to some classifiers [23].

Some of previous work listed in [4, 33, 58-60] as an examples for rank level fusion with fusion approaches used and modalities fused. In general, it remains significantly understudied.

3) Decision level fusion

Principles- The final decision - in multimodal biometric systems - is formed from obtaining individually separate decision of different biometric modalities using different techniques include behavior knowledge space, majority voting, , weighted voting, AND rule, and OR rule[5, 8]. Decision level fusion is also named abstract level fusion; because it is used when there is access to only decisions from individual [8, 23].

Majority voting approach is the mostly used for decision level fusion. The input sample with agreed in majority of matchers is given the identity. AND/OR rules are rarely used; because they combine two different matchers, so this sometimes degrade of performance of the system. AND combination improves the FAR while, OR combination improves the FRR. The main advantage of the majority voting method is that it does not require prior knowledge about the matcher, and it requires no training for final decision making too [42].

Discussion- Decision level fusion approaches are well investigated for biometric systems but are too rigid (inflexible) because of availability of limited amount information; probability of having a tie may appear [4]. And only consider single information for fusion, which has a high probability of producing wrong recognition result [5, 18]. As it have a less amount of features or scores information of different modalities; it is very easy to implement. [23]. This type is less preferred in multi-biometric system implementation.

Decision level fusion based examples include: majority voting rule and behavioral knowledge space method, weighted voting based on Dempster - Shafer theory, AND/ OR rules for deciding the decision , and that naïve Bayesian decision fusion as it works well, even if the matchers used in fusion are dependent to each other. In addition, some of other last research found in [61-64].

C. Hybrid Level Fusion

Tri-level fusion scenarios (different fusion in different levels of the system) can be investigated to make the system faster and significantly reduce the error rate. The fusion of level increased the performance. In 2007, C. Lupu et al. [65] fused fingerprint, voice and iris. Next year 2008, S. Asha et al. [7] combined fingerprint with mouse dynamics. In 2011, Parallel Feature Extraction with the help of SIFT, SIMD, and HMA techniques was used by Anukul Chandra Panda et al.[66] to fuse multiple iris. Next in 2013, Gandhimathi Amirthalingam, and Radhamani. G. [5] used fuzzy vault to implement multimodal system based on Face and ear traits. Some examples of previous work used such fusions are in [42, 67-71]. Fusion approaches, fusion levels, and performance for these papers ordered by year, are listed in table (II) below.

VII. Design and Implementation of Multi-Biometrics Recognition Trade-Offs

Generally, any biometric recognition system architecture is related to software-based techniques and hardware-based techniques. The obstacles here is to satisfy all challenges requirement such as: user friendly, fast (i.e. the system must identify individuals in real time), low cost, high performance, less intrusive, fraud prevent and high fake detection rate [72]. Briefly, design issues in multi-biometrics include [17]:

- Choosing the biometric modalities and number of traits (defining and estimation of each modality reliability is still open research issue).
- Choosing the best samples for a particular biometric.
- Fusion level and fusion methodology.
- Fusion scenario and common strategy.
- Learning weights of individual biometric for users.
- Cost versus performance and accuracy versus reliability trade-offs.
- Verification and/or identification system for application.
- Expert features selection difficulties.

In order to optimize the multi-biometric recognition benefits, the issues of system design firstly should be understood better; so the more effective design methodology and system architecture can be developed. For instance, to decide whether combining multiple biometrics or combining multiple samples of the same trait is better, to achieve economic system. In addition, privacy issues should be considered, and compromising between accuracy and coverage.

VIII. Multi-Biometrics - Discussion and Research Direction

Several research directions arise from the work proposed in this topic. There are some issues and open questions still need some efforts. We suggest the following tasks and discussion as future work that would significantly improve the security or other performance metrics of multi-biometric systems. Below is a hot point in this field still under research.

A. Multi-data Database / Real dataset

A dataset is not a research result in itself but, a well-designed one can facilitate the research. Many researchers are putting efforts in fusing multimodal biometrics. There are different approaches for biometric fusion. One approach is to use heterogeneous database (i.e. one biometric trait from one database and other trait from another database). But this approach is not reflecting the performance of multimodal users. The other approach, is to use homologous database. It means different biometrics from the same person. Only few multimodal databases are available publicly [73]. BANCA and XM2VTS includes face and voice biometrics. BIOMET which includes face, voice, fingerprint, hand and signature. BIOSEC includes fingerprint, ace, iris and voice. SDUMLAHMT is a homologous database which includes face images from 7 angles, finger print images, gait videos, iris images. But these databases have some limitations. Homologous multi-biometrics dataset should be complete (contains all the biometrics for large population) for future research testing and multi-biometric system evaluation.

B. *Soft Multi-biometrics*

Using multiple biometric identifiers in a single system will increase the identification or verification times and hence, cause more inconvenience to the users and increase the overall cost of the system. Thus, soft biometric is introduced in 2004 to obtain the same recognition performance without causing any additional inconveniences to the users by incorporating it (soft biometric identifiers) to the primary multimodal systems [8]. Soft biometric identifiers include gender, ethnicity, height, weight, eye color, skin color, hair color, etc. Two key challenges need to be addressed to incorporate soft biometrics into the traditional multimodal biometric framework. The first challenge, is the automatic and reliable extraction of the soft biometric information without causing inconveniences to the users, and the second challenge, is to combine optimally this information with the primary biometric identifier to achieve the best recognition performance. Soft multi-biometrics could be implemented by using Oracle or SQL Server programming language tool that integrates the database implementation with pattern recognition and image processing techniques.

C. *Multi-Algorithms fusion methods*

Such systems seek to improve the speed, reliability, and accuracy of a biometric system. A variety of fusion methods and approaches have been described in [14]. We suggest new methods and modified algorithms to build and test the multi-biometric system. In [56], a new robust linear programming method proposed theoretically to fuse multi-biometrics by combining the modalities optimally. The robustness and accuracy have to be practically measured.

Another suggestion is to adopt K-means to cluster data and other advanced clustering methods to offer the best solutions especially when data are influenced by kinds of noise. The new modified feature descriptor Scale Invariant Feature Transform (F-SIFT) algorithm, Incremental Granular Relevance Vector Machine (iGRVM), Particle Swarm Optimization (PSO), and Hidden Markov Models (HMM) have not been used practically yet as new fusion techniques. The performance of multi-unit biometric trait recognition may be improved. Also, using the classifiers in matching fusion is still under research. In the multimodal biometric literature, a lot of attention has been paid to the parallel fusion of multiple classifiers. A few of reported works dealt so far with serial architecture. It would also be of interest to study the performance of the proposed techniques with the serial fusion of multiple classifiers using F-SIFT, iGRNM, PSO, and HMM algorithms suggested.

D. *Identification of Identical Twins*

The identification of identical twins is a big challenge, as the unimodal system is less accuracy in this state. Twins are the most similar persons in terms of genetics. The multimodal can increase the recognition rate as the Twins cannot have the same modalities together. Face, fingerprint, and iris could be fused to identify twins. To extend the study on the similarity of biometrics of identical twins, the use of siblings' data would be a hot point in future.

E. *Indexing Search (Time and Complexity Enhancement)*

During identification mode, search time plays a significant role. The search space of large biometrics database can be reduced through indexing and cloud computing. Various local feature based indexing approaches are proposed using multi-dimensional trees. Though k-d tree improves searching time, but insertion into the tree is not dynamic [54]. This is not suitable as databases are continuously updated to new enrollments. Another data structure known as k-d-b tree suggested to resolve such these issues. To improve the rank of identification for R-tree indexing, a hybrid coarse-to-fine searching strategy will be proposed. Also, we suggest parallel sorting of vote counts using Hypercube Mesh Architecture (HMA) in order to retrieve the image and get the top k matches; this may achieve less in time and complexity, when indexing scores are combined with match scores. Indexing using parallel geometric hashing is faster and could find its applicability in various real-time applications. All these points, if practiced upon multi-biometrics over cloud computing topology, it may become a solution for some biometric architecture design issues. Some problems and promises of using the cloud and biometrics are discussed in [74].

F. *Embedded Hybrid Recognition System*

From the above survey, some points noticed as a few research used sensor level fusion; we suggest fusion between physiological and behavioral traits such (iris, fingerprint, face…etc.) with (gait, signature). Fusion between the offline and online signature acts more authentication for critical documents signing. At the same time, the multi-fusion also can be used with multi classifiers and using different fusion levels. The multi-biometric system then may be more complex. This can be resolved by using the parallelism in feature extraction and identification phases, or execution by using H/W devices like Arduino or FPGA or parallel processing elements. In most cases, multi-biometric based security systems need to operate actively in the real-time public network and authentication environment.

IX. CONCLUSION

Multi-biometrics topic has attracted more interest in recent research. It is used to identify individuals based on their physiological and behavioral characteristics for security purposes. Overview of biometrics showed that it is impossible to find the best single biometric suitable for all applications, populations, technologies and administration policies. Also, integration of biometric modalities can solve unimodal system limitations to achieve higher performance.

Benefits and limitations of multi-biometrics discussed as we introduced it as a solution. In this paper, a state of the art survey of integration strategies, and fusion levels prior to matching and after matching are discussed with advantages and disadvantages of each type. However, Design and evaluate the multi-biometric systems raises many issues and trends. Finally, some open points suggested to be considered as a future research and enhance applications.

TABLE II. SOME UPTODATE EXAMPLES OF PREVIOUS RESEARCH BASED DIFFERENT FUSION IN DIFFERENT LEVELS

Year	Modalities fused	Author(s)	Fusion Level	Fusion Approach	Performance in percentage
2004	Fingerprint + Face	Kalyan, et al.[67]	Score + Decision	Sum Rule and Likelihoods	58.33% improvement with correlation 0.9 And (sum rule, PSO)=(0.0324,0.0135)%
2011	Face + Palm print	Linin Shen [68]	Feature+ Decision	FPCODE	Feature level fusion : 91.52% Decision level fusion : 91.63%
2013	Face + Ear	S.M.S. Islam[69]	Feature + Score	L3DF, Iterative closet point	FAR = 0.001 % Recognition: 96.8% Verification: 97.1%
2014	Face + Fingerprint + Iris	A. Annis Fathima et al. [71]	Score + Dynamic decision	Weighted average fusion, and K-NN	Recognition Rate= 78.5484% (Iris + Face) = 85%

REFERENCES

[1] M. Manjunath and K. B. Raja, "A Novel Approach for Iris Recognition using DWT&PCA," Int. J. Advanced Networking and Applications, vol. 5, no. 1, pp. 1830-1836, 2013.

[2] S. Malhotra and D. C. Kant, "A Novel Approach for Securing Biometric Template," Internal Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 3, no. 5, pp. 397-403, 2013.

[3] A. Bhargava and R. S. Ochawar, "Biometrics in Access Control System," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 2, pp. 269-273, 2013.

[4] N. Radha and A. Kavitha, "Rank Level Fusion Using Fingerprint and Iris Biometrics," Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 6, pp. 917-923, 2011.

[5] G. Amirthalingam, "A Multimodal Approach for Face and Ear Biometric System," International Journal of Computer Science Issues (IJCSI), vol. 10, no. 5, pp. 234-241, 2013.

[6] S. Sharma, "An Improved Iris Recognition System Based on 2-D DCT and Hamming Distance Technique," ICRTEDC-2014, GV/ICRTEDC/08, vol. 1, no. 2, pp. 32-34, 2014.

[7] S. Asha and C. Chellappan, "Authentication of E-Learners Using Multimodal Biometric Technology," presented at the Biometrics and Security Technologies, 2008. ISBAST 2008. International Symposium on, Islamabad, 2008. pp. 1-6.

[8] A. A. Ross, K. Nandakumar, and A. K. Jain, Handbook of Multibiometrics vol. 6. New York: Springer Science & Business Media, 2006.

[9] M. L. Gavrilova and M. M. Monwar, "Current Trends in Multimodal Biometric System Rank Level Fusion," in Pattern Recognition, Machine Intelligence and Biometrics, ed: Springer, 2011, pp. 657-673.

[10] R. Singhal, N. Singh, and P. Jain, "Towards An Integrated Biometric Technique," International Journal of Computer Applications, vol. 42, no. 13, pp. 20-23, 2012.

[11] Y. S. a. S. Singh, "Evaluation of Electrocardiogram for Biometric Authentication," Journal of Information Security, vol. 3, no. 1, pp. 39-48, 2012.

[12] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 4-20, 2004.

[13] F. Karray, J. A. Saleh, M. N. Arab, and M. Alemzadeh, "Multi Modal Biometric Systems: A State of the Art Survey," Pattern Analysis and Machine Intelligence Laboratory, University of Waterloo, Waterloo, Canada, no. 2007.

[14] A. M. Siddiqui, R. Telgad, and P. D. Deshmukh, "Multimodal Biometric Systems: Study to Improve Accuracy and Performance," International Journal of Current Engineering and Technology, vol. 4, no. 1, pp. 165-171, 2014.

[15] H. AlMahafzah and M. Z. AlRwashdeh, "A Survey of Multibiometric Systems," International Journal of Computer Applications vol. 43, no. 15, pp. 36-43, 2012.

[16] K. Delac and M. Grgic, "A Survey of Biometric Recognition Methods," in Electronics in Marine, 2004. Proceedings Elmar 2004. 46th International Symposium, ELMAR-2004, Zadar, Croatia, 2004, pp. 184-193.

[17] M. S. Ahuja and S. Chabbra, "A Survey of Multimodal Biometrics," International Journal of Computer Science and its Applications, vol. 1, no. pp. 157-160, 2011.

[18] A. A. Ross, A. K. Jain, and K. Nandakumar, "Information Fusion in Biometrics," in Handbook of Multibiometrics, ed, 2006, pp. 37-58.

[19] G. Feng, K. Dong, D. Hu, and D. Zhang, "When Faces are Combined With Palmprints: A novel Biometric Fusion Strategy," presented at the In proc. of 1st Int. Conf. on Biometric authentication, Hong Kong, China, 2004. pp. 701-707.

[20] R. Gad, M. Mohamed, and N. El-Fishawy, "Iris Recognition Based on Log-Gabor and Discrete Cosine Transform Coding," Journal of Computer Science and Engineering, vol. 5, no. 2, pp. 19-26, 2011.

[21] R. Bhatia, "Biometrics and Face Recognition Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 5, pp. 93-99, 2013.

[22] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of Artificial Gummy Fingers on Fingerprint Systems," in Electronic Imaging 2002, Proceedings of SPIE, San Joes, USA, 2002, pp. 275-289.

[23] D. T. Meva and C. K. Kumbharana, "Comparative Study of Different fusion techniques in multimodal biometric authentication," International Journal of Computer Applications, vol. 66, no. 19, 2013.

[24] S. Kalra and A. Lamba, "A Survey on Multimodal Biometric," International journal of computer science and information technologies, vol. 5, no. 2, pp. 2148-2151, 2014.

[25] M. A. P. C. Mr. Rupesh Wagh, "Analysis of Mutlimodal Biometrics with Security Key," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE). vol. 3, no. 8, pp. 1363-1365, 2013.

[26] A. Meraoumia, S. Chitroub, and A. Bouridane, "Multimodal Biometric Person Recognition System based on Iris and Palmprint Using Correlation Filter Classifier," in Proc. of the Second International Conference on Communications and Information Technology, Hammamet, Tunisia, June 26-28, 2012, pp. 782-787.

[27] M. Deriche, "Trends and Challenges in Mono and Multi biometrics," presented at the Image Processing Theory, Tools and Applications, 2008. IPTA 2008. First Workshops on, Sousse, 2008. pp. 1-9.

[28] N. Geethanjali and K. Thamaraiselvi, "Feature Level Fusion of Multimodal Biometrics and Two Tier Security in ATM System," International Journal of Computer Applications, vol. 70, no. 14, pp. 17-23, 2013.

[29] D. Satyarthi, Y. P. S. Maravi, P. Sharma, and R. K. Gupta, "Comparative Study of Offline Signature Verification Techniques," International Journal of Advancements in Research & Technology, vol. 2, no. 2, pp. 1-6, 2013.

[30] G. Sathish, S. V. Saravanan, S. Narmadha, and S. U. Maheswari, "Multi-Algorithmic Iris Recognition," International Journal of Computer Applications, vol. 38, no. 11, pp. 13-21, 2012.

[31] K. Elumalai and M. Kannan, "Multimodal Authentication for High End Security," International Journal on Computer Science and Engineering, vol. 3, no. 2, pp. 687-692, 2011.

[32] B. Schouten and B. Jacobs, "Biometrics and Their Use in E-Passports," Image and Vision Computing, vol. 27, no. 3, pp. 305-312, 2009.

[33] M. Monwar and M. L. Gavrilova, "Multimodal Biometric System Using Rank-Level Fusion Approach," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,, vol. 39, no. 4, pp. 867-878, 2009.

[34] M. Soltane, N. Doghmane, and N. Guersi, "Face and Speech Based Multi-Modal Biometric Authentication," International Journal of Advanced Science and Technology, vol. 21, no. 6, pp. 41-56, 2010.

[35] F. Anwar, M. A. Rahman, and S. Azad, "Multibiometric Systems Based Verification Technique," European J. Scientific Research, vol. 34, no. 2, pp. 260-270, 2009.

[36] R. S. Choras, "Hybrid Iris and Retina Recognition for Biometrics," presented at the Image and Signal Processing (CISP), 2010 3rd International Congress on, Yantai,, 2010. pp. 2422-2426.

[37] Z. Huang, Y. Liu, C. Li, M. Yang, and L. Chen, "A Robust Face and Ear Based Multimodal Biometric System Using Sparse Representation," Pattern Recognition, vol. 46, no. 8, pp. 2156-2168, 2013.

[38] L. Lu and J. Peng, "Finger Multi-biometric Cryptosystem using Feature-Level Fusion," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 7, no. 3, pp. 223-236, 2014.

[39] A. S. Makinde, Y. Nkansah-Gyekye, and L. S. Laizer, "Enhancing the Accuracy of Biometric Feature Extraction Fusion Using Gabor Filter and Mahalanobis Distance Algorithm," (IJCSIS) International Journal of Computer Science and Information Security, vol. 12, no. 7, pp. 1-8, 2014.

[40] P. A. KUMARI and G. J. SUMA, "A Novel Mutimodal Biometric Scheme or Personal Authentication," IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET) vol. 2, no. 2, pp. 55-66 2014.

[41] S. Kalra and A. Lamba, "Improving Performance by Combining Fingerprint and Iris in Multimodal Biometric," International Journal of Computer Science & Information Technologies, vol. 5, no. 3, pp. 4522-4525, 2014.

[42] A. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Biometric Systems," Pattern Recognition, vol. 38, no. 12, pp. 2270-2285, 2005.

[43] A. R. a. R. Govindarajanb, "Feature Level Fusion Using Hand and Face Biometrics," presented at the SPIE Conference on Biometric Technology for Human Identification II, Orlando, USA, 2005. pp. 196-204.

[44] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood Ratio-Based Biometric Score Fusion," IEEE Transactions on Pattern Analysis and Machine Intelligence,, vol. 30, no. 2, pp. 342-347, 2008.

[45] M. Nageshkumar, P. Mahesh, and M. S. Swamy, "An Efficient Secure Multimodal Biometric Fusion Using Palmprint and Face Image," (IJCSI) International Journal of Computer Science Issues, vol. 2, no. 4, pp. 49-53, 2009.

[46] A. Darwish, R. A. Elghafar, and A. F. Ali, "Multimodal Face and Ear Images," Journal of Computer Science, vol. 5, no. 5, p. 374, 2009.

[47] M. I. Razzak, R. Yusof, and M. Khalid, "Multimodal Face and Finger Veins Biometric Authentication," Scientific Research and Essays, vol. 5, no. 17, pp. 2529-2534, 2010.

[48] P. Dalka and A. Czyzewski, "Human-Computer Interface Based on Visual Lip Movement and Gesture Recognition," IJCSA, vol. 7, no. 3, pp. 124-139, 2010.

[49] M. Kawulok and J. Szymanek, "Precise Multi-Level Face Detector for Advanced Analysis of Facial Images," Image Processing, IET, vol. 6, no. 2, pp. 95-103, 2012.

[50] F. Alsaade, "Neuro-Fuzzy Logic Decision in a Multimodal Biometrics Fusion System," Scientific Journal of King Faisal University (Basic and Applied Sciences), vol. 11, no. 2, p. 14, 2010.

[51] F. CUI and G. Yang, "Score Level Fusion of Fingerprint and Finger Vein Recognition," Journal of Computational Information Systems, vol. 7, no. 16, pp. 5723-5731, 2011.

[52] M. Romaissaa and R. Abdellatif, "On Comparing Verification Performances of Multimodal Biometrics Fusion Techniques," International Journal of Computer Applications, vol. 33, no. 7, pp. 24-29, 2011.

[53] A. A. Paulino, "Contributions to Biometric Recognition: Matching Identical Twins and Latent Fingerprints," PhD degree of Computer Science, Michigan State University, 2013.

[54] H. Mehrotra, "On the Performance Improvement of Iris Biometric System," PhD, Departmant of Computer Science and Engineering, National Institute of technology Rourkela, Rourkela, Odisha, India, 2014.

[55] H. M. Sim, H. Asmuni, R. Hassan, and R. M. Othman, "Multimodal Biometrics: Weighted Score Level Fusion Based on Non-Ideal Iris and Face Images," Expert Systems with Applications, vol. 41, no. 11, pp. 5390-5404, 2014.

[56] D. Miao, Z. Sun, and Y. Huang, "Fusion of Multibiometrics Based on A New Robust Linear Programming," presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on, Stockholm, 2014. pp. 291-296.

[57] A. Gambhir, S. Narke, S. Borhade, and G. Bokade, "Person Recognition Using Multimodal Biometrics," International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol. 4, no. 4, pp. 725-728, 2014.

[58] M. Monwar and M. Gavrilova, "Secured Access Control Through Markov Chain Based Rank Level Fusion Method," presented at the in proc. of 5th Int. Conf. on Computer Vision Theory and Applications (VISAPP), Angres, France, 2010. pp. 458-463.

[59] A. Kumar and S. Shekhar, "Personal Identification Using Multibiometrics Rank-Level Fusion," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, , vol. 41, no. 5, pp. 743-752, 2011.

[60] M. MONWAR, "A Multimodal Biometric System Based on Rank Level Fusion," PhD, Department of Computer Science University of Calgary, ALBERTA 2012.

[61] S. A. S. DzatiAthiarRamli, AiniHussain, "A Multibiometric Speaker Authentication System with SVM Audio Reliability Indicator," International Journal of Computer Science & Information Technologies (IAENG), vol. 36, no. 4, pp. 313-321, 2008.

[62] S. K. Grewal, "A Composite Approach for Biometric Template Security," International Journal and Conference Service Center (IJCSC), vol. 5, no. 1, pp. 170-176, 2014.

[63] I. A. Saleh and L. M. Alzoubiady, "Decision Level Fusion of Iris and Signature Biometrics for Personal Identification using Ant Colony Optimization," International Journal of Engineering and Innovative Technology (IJEIT), vol. 3, no. 11, pp. 35-42, 2014.

[64] A. Naghate, M. Sahu, P. Bhange, S. Lonkar, P. Wankhede, and Y. Bute, "Implementation of Multibiometric System Using Iris and Thumb Recognition," International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 3, no. 3, pp. 932 – 940, 2014.

[65] C. Lupu and V. Lupu, "Multimodal Biometrics for Access Control in An Intelligent Car," presented at the Computational Intelligence and Intelligent Informatics, 2007. ISCIII'07. International Symposium on, Agadir, 2007. pp. 261-267.

[66] A. C. Panda, "Parallel Algorithms for Iris Biometrics," M.Sc., Department of Computer Science and Engineering, National Institute of Technology Rourkela, Odisha, India, 2011.

[67] K. Veeramachaneni, L. Osadciw, A. Ross, and N. Srinivas, "Decision-Level Fusion Strategies for Correlated Biometric Classifiers," presented at the Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, Anchorage, AK 2008. pp. 1-6.

[68] L. Shen, L. Bai, and Z. Ji, "FPCODE: An Efficient Approach for Multi-Modal Biometrics," International Journal of Pattern Recognition and Artificial Intelligence, vol. 25, no. 02, pp. 273-286, 2011.

[69] S. M. Islam, R. Davies, M. Bennamoun, R. A. Owens, and A. S. Mian, "Multibiometric Human Recognition Using 3D Ear and Face Features," Pattern Recognition, vol. 46, no. 3, pp. 613-627, 2013.

[70] W. Almayyan, "Performance Analysis of Multimodal Biometric Fusion," PhD, Faculty of Technology, De Montfort University, England, United Kingdom, 2012.

[71] A. A. Fathima, S. Vasuhi, N. N. Babu, V. Vaidehi, and T. M. Treesa, "Fusion Framework for Multimodal Biometric Person Authentication System," IAENG International Journal of Computer Science, vol. 41, no. 1, pp. 1-14, 2014.

[72] J. Galbally, S. Marcel, and J. Fierrez, "Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint, and Face Recognition," IEEE Transactions on Image Processing, , vol. 23, no. 2, pp. 710-724, 2014.

[73] V. SIREESHA and K. SANDHYARANI, "Multimodal Biometric System Using Iris-Fingerprint: An Overview," International Journal of Engineering Sciences Research-IJESR, vol. 02, no. Special Issue 01, pp. 1342-1349, 2013.

[74] A. A. Albahdal and T. E. Boult, "Problems and Promises of Using the Cloud and Biometrics," presented at the 11th International Conference on Information Technology: New Generations (ITNG)2014,, 2014. pp. 293-300.

Application of Image Processing Techniques for TV Broadcasting of Sporting Events

Cheikhrouhu E.
The National Higher Engineering School of Tunis (ENSIT,
University of Tunis, L.R: LATICE, Tunisia

Mlouhi Y.
The National Higher Engineering School of Tunis (ENSIT,
University of Tunis, L.R: LATICE, Tunisia

Jabri I.
The National Higher Engineering School of Tunis (ENSIT,
University of Tunis, L.R: LATICE, Tunisia

Battikh T.
The National Higher Engineering School of Tunis (ENSIT,
University of Tunis, L.R: LATICE, Tunisia

Lakhoua M.N.
The National School of Engineering of Carthage, University
of Carthage, U.R: SMS, Tunisia

Maalej L.
The National Higher Engineering School of Tunis (ENSIT,
University of Tunis, L.R: LATICE, Tunisia

Abstract—**In this paper, we describe a system solution thanks to which virtual graphics, the projection of advertising images, logos, match scores and the distance measurements of players on the field may be overlaid on the plan of the different types of sports fields of real tested images. This solution relies on a study of the artificial vision and the Augmented Reality applied to TV broadcasting of sporting events where we have as input the original image to be processed, the image to be projected and the coordinates of the overlay position of the objects on the plan of the field. As an output, we have the overlaid objects in the processed image at the selected position in a more realistic way and in the background.**

Keywords—Augmented Reality; Filtering; Sports field Homography; Image Processing

I. Introduction

Augmented reality is one of the fields of artificial or computer-assisted vision. It is a rapidly growing research area thanks to its underlying principle allowing the mixing of the real and virtual worlds [1], [2].

Thus, the objective of the research work carried out in the field of augmented reality has so far been primarily focused on the positioning in real time of virtual objects in a real scene [3].

To this end, appropriate measurement tools are needed. So is an adequate processing environment dedicated for the processing of images and information [4]. The work we describe in this article purports to provide a solution which would enable to overlay graphics on the plan of a soccer pitch, tennis, handball or basketball courts, by embedding for instance, adverts, flags, match scores or the analysis of the match or the measurements of the distance of the players in relation to the goal on the field.

To do so, we need to make sure that the insertion of the objects on the field occurs in the most real way and under the objects in the foreground (players, tennis net, etc.) The combination of methods used has enabled us to reach a high level of reliability and robustness as far as the obtained results on the tested real images are concerned.

II. Review on Atificial Vision and Augmented Reality

We present some studies on the artificial vision and augmented reality that have been presented in various researches:

Researchers Seong-Oh & al. [5], have proposed a new mobile augmented-reality system that will address the need of users in viewing baseball games with enhanced contents. The overall goal of the system is to augment meaningful information on each player position on a mobile device display. To this end, the system takes two main steps which are homography estimation and automatic player detection. This system is based on still images taken by mobile phone. The system can handle various images that are taken from different angles with a large variation in size and pose of players and the playground, and different lighting conditions. They have implemented the system on a mobile platform. The whole steps are processed within two seconds.

Researchers Stricker, D. & al. [6], have presented several results from the research department "Augmented Vision" of the German Research Center for Artificial Intelligence. The driving idea of this work is to move from traditional Augmented Reality (AR) systems, which are often limited to visualization and tracking components, to AR cognitive systems, which have or gradually build knowledge about the situation and intentions of the user. Such systems will basically be much more unobtrusive and adapt the information presentation to the users' actual needs. To reach this goal, strong progress must be done in several areas, starting with 3D scene digitalization and analysis, body modeling and motion capturing, and action and workflow recognition. An overview of current results and work-in-progress of the Augmented Vision group in those areas is presented and finally discussed.

Researchers Rui & al. [7], have discussed the vision-based registration techniques for augmented reality (AR) systems which have been the subject of intensive research recently due to their potential to accurately align virtual objects with the real world. The downfall of these vision-based approaches, however, is their high computational cost and lack of robustness. To address these shortcomings, a robust pose estimation algorithm based on artificial planar markers is adopted. This algorithm solves the problem of camera pose ambiguities and is able to draw a unique and robust solution. Experiments show the robustness and effectiveness of this method in the context of real-time AR tracking.

Researchers Lourakis & al. [8], have presented in their work the camera matchmoving witch is an application involving synthesis of real scenes and artificial objects, in which the goal is to insert computer-generated graphical 3D objects into live-action footage depicting unmodeled, arbitrary scenes. This work addresses the problem of tracking the 3D motion of a camera in space, using only the images it acquires while moving freely in unmodeled, arbitrary environments. A novel feature-based method for camera tracking has been developed, intended to facilitate tracking in online, time-critical applications such as video see-through augmented reality and vision-based control. In contrast to several existing techniques, which are designed to operate in a batch, offline mode, assuming that the whole video sequence to be tracked is available before tracking commences, the proposed method operates on images incrementally, as they are being acquired.

Researchers Ji Hoon Choi & al. [9], have introduced a method for personalized data broadcasting service using TVA metadata. Appropriate metadata structure for personalized data broadcasting is explained by comparison with package metadata and data broadcasting contents. New scenario and contents structure for personalized data broadcast is described. Therefore the system and data flow mechanism for personalized data broadcasting service is presented. In fact, the number of broadcasting channels and contents are increasing with the arrival of digital broadcast and various broadcasting medium. However, there is a limit on searching of the program by using conventional program guide.

III. Presentation of the Proposed Solution

The implementation process of our solution consists mainly of three stages (See Figure 1):

- Pre-processing which aims at filtering the colors in the background of the image.
- Processing which combines the computational methods of the Homography matrix corresponding to the image selected and the insertion of the graphics projected on the field plan.
- Post-processing which presents the display of the outcome image.

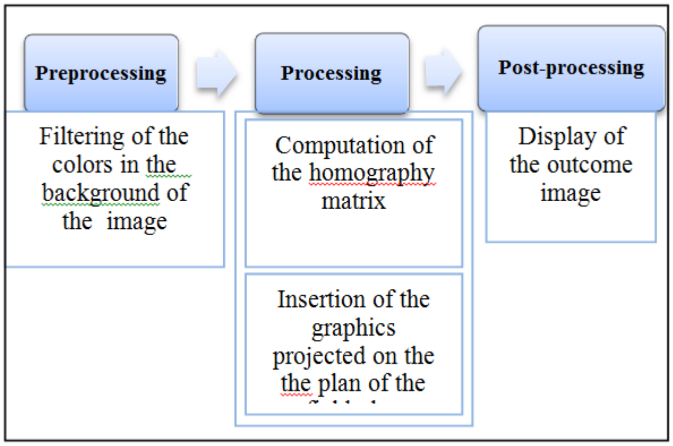

Preprocessing
Processing
Post-processing
Filtering of the colors in the background of the image
Computation of the homography matrix
Insertion of the graphics projected on the the plan of the
Display of the outcome image

Fig. 1. Structure of the proposed solution

A. Stage 1: Background color filtering method

This method consists in eliminating the background color of the field of the selected image. some background colors should just be selected to use them later to carry out the filtering whose aim is to detect, for instance, the players on the field and the net (in the case of a tennis court) and all the objects that are in the foreground [10].

The Algorithm Background color Filtering Method:

```
0) Start
1) Select the position of the colors to use for filtering
   (from one to four entrance points) the background of
   the field of the selected image.
2) For  i to 1 over the width of the selected image.
    For j to 1 over the length of the selected image
      If the pixel color of the position (i, j) of the image is
      close to one the colors selected at the beginning
      Then it is not to be put into the filtered image
      If not
      Put the color of this pixel at the position (i, j) in the
      filtered image.
      End If
      End For
      End For
3) Display and save the filtered image to use it later
4) End
```

B. Stage 2: Computational method of the Homography matrix

The computation of the Homography matrix is the second stage in our solution thanks to which we will be able to calculate the matrix of the perspective correction and the liaison between the selected image and the corresponding field image, depending on the type of field (tennis, soccer, Handball, Basketball) so that we can use it later during the overlaying of the objects on the plan of the field which would represent the overall perspective projection matrix of the four points of the field and their four corresponding points on the original image of the selected field.

When plan objects project themselves on a captor, the images seen from different perspective points (straight or oblique) are linked through a projective transformation called " homography" of the form: P' = H P with each point P(x, y) having its corresponding P'(x', y').

Homography «H» is characterized by a 3x3 homogeneous matrix. This transformation induces a scale factor and includes only 8 independent coefficients (h33=1) [2].

With w and w' being the homogeneous coordinates.

To adjust an image, we need to determine the 8 coefficients of the H transformation which brings the points of the plane source image to a benchmark position. The points P and P' are linked by:

$$\begin{pmatrix} x' \\ y' \\ w' \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x \\ y \\ w \end{pmatrix}$$

$$x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}} \qquad y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}$$

With the 4 points: P0(x0,y0), P1(x1,y1), P2(x2,y2) et P3(x3,y3) of the source image being assigned to positions P'0(x'0,y'0), P'1(x'1,y'1), P'2(x'2,y'2) and benchmark P'3(x '3,y'3), W have de 8 equations with 8 unknowns :

$$x'0\ (h_{31}x0 + h_{32}y0 + h_{33}) = h_{11}x0 + h1_{2}y0 + h_{13}$$

$$y'0(h_{31}x0 + h_{32}y0 + h_{33}) = h_{21}x0 + h_{22}y0 + h_{23}$$

$$x'1\ (h_{31}x1 + h_{32}y1 + h_{33}) = h_{11}x1 + h_{12}y1 + h_{13}$$

$$y'1(h_{31}x1 + h_{32}y1 + h_{33}) = h_{21}x + h_{22}y1 + h_{23}$$

$$x'2\ (h_{31}x2 + h_{32}y2 + h_{33}) = h_{11}x2 + h_{12}y2 + h_{13}$$

$$y'2(h_{31}x2 + h_{32}y2 + h_{33}) = h_{21}x2 + h_{22}y2 + h_{23}$$

$$x'3\ (h_{31}x3 + h_{32}y3 + h_{33}) = h_{11}x3 + h_{12}y3 + h_{13}$$

$$y'3(h_{31}x3 + h_{32}y3 + h_{33}) = h_{21}x3 + h_{22}y3 + h_{23}$$

Hence, with these 4 control points, we obtain the following system:

$$\underbrace{\begin{pmatrix} x0 & y0 & 1 & 0 & 0 & 0 & -x0\,x'0 & -y0\,x'0 \\ 0 & 0 & 0 & x0 & y0 & 1 & -x0\,y'0 & -y0\,y'0 \\ x1 & y1 & 1 & 0 & 0 & 0 & -x1\,x'1 & -y1\,x'1 \\ 0 & 0 & 0 & x1 & y1 & 1 & -x1\,y'1 & -y1\,y'1 \\ x2 & y2 & 1 & 0 & 0 & 0 & -x2\,x'2 & -y2\,x'2 \\ 0 & 0 & 0 & x2 & y2 & 1 & -x2\,y'2 & -y2\,y'2 \\ x3 & y3 & 1 & 0 & 0 & 0 & -x3\,x'3 & -y3\,x'3 \\ 0 & 0 & 0 & x3 & y3 & 1 & -x3\,y'3 & -y3\,y'3 \end{pmatrix}}_{\mathbf{T}} * \underbrace{\begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{pmatrix}}_{\textit{The coefficients of « H »}} = \underbrace{\begin{pmatrix} x'0 \\ y'0 \\ x'1 \\ y'1 \\ x'2 \\ y'2 \\ x'3 \\ y'3 \end{pmatrix}}_{\textit{The Points' coordinates}}$$

The resolution of this linear system will enable us to calculate the matrix of the H perspective remedial Homography as well as these 8 coefficients [11].

Algorithm of the homography matrix computation method :

0) Start

1) Select the type of field of the selected image

2) For i from 0 to 3 (with i being the chosen point index ranging from 0 to 3)

- Recording of the (xi,yi) coordinates on the side of the selected Pi point in the «coor_cot» matrix of the selected image coordinates such that coor_cot(i,0) = xi and coor_cot(i,1) = yi.

  Recording of the (x'i,y'i) side coordinates of the corresponding P'i point in the «coor_originale_cot» matrix of the original coordinates of field image such that coor_originale_cot(i,0) = x'i and coor_originale_cot(i,1) = y'i.

End for

3) Transformation of the «coor_cot» and «coor_originale_cot» matrices into 2 «cord» and «cordorign» tables of size equal to 8 :

- cord = [coor_cot (0,0), coor_cot (0,1),
- coor_cot (1,0), coor_cot (1,1), coor_cot (2,0), coor_cot (2,1), coor_cot (3,0), coor_cot (3,1)]
- cordorign = [coor_originale_cot(0,0),

- coor_originale_cot(0,1), coor_originale_cot(1,0), coor_originale_cot(1,1), coor_originale_cot(2,0), coor_originale_cot(2,1), coor_originale_cot(3,0), coor_originale_cot(3,1)]

4) The filling-in of the 8*8 “T” Matrix according to the method [3] from the coordinates of the selected « cord» and «cordorign» tables in order to use them later in the computation of the Homography «H» matrix coefficients :

- For j From 0 to 6 (not = 2) (with j being the index of Table T rows)

```
    T(j, 0) = cord (j)
    T(j, 1) = cord (j+1)
    T(j, 2) = 1
    T(j, 3) = 0
    T(j, 4) = 0
    T(j, 5) = 0
    T(j, 6) = -1 * cord (j) * cordorign (j)
    T(j, 7) = -1 * cord (j+1) * cordorign (j)
    T(j + 1, 0) = 0
    T(j + 1, 1) = 0
    T(j + 1, 2) = 0
    T(j + 1, 3) = cord (j)
    T(j + 1, 4) = cord (j+1)
    T(j + 1, 5) = 1
    T(j + 1,6) = -1 * cord (j) * cordorign (j+1)
    T(j + 1, 7) = -1 * cord (j+1) * cordorign (j+1)
```

- End for

5) Computation of «T^{-1}» matrix which represents the inverse matrix of the «T» matrix and then multiply it by the «coor_originale_cot» table so as to eventually have the 8 coefficients of the «H» matrix and fill them in an «Hcoef» size 8 table (being a product of a matrix and a vector):

- For l from 0 to 7
- For c from 0 to 7
- Hcoef(l) = Hcoef(l) + (cordorign (l) * T^{-1}(l, c))
- End for
- End for

6) The filling-in of the final Homography matrix “H” of a 3*3 size with the 8 computed coefficients with H(3,3)=1 and display it afterwards :

- H= {[Hcoef(0), Hcoef(1), Hcoef(2)],
- [Hcoef(3), Hcoef(4), Hcoef(5)],
- [Hcoef(6), Hcoef(7), 1]}

7) End

C. Stage 3: Method of overlaying graphics on the field’s plan

This method makes it possible to superimpose graphics on the field’s plan in accordance with a selected position and according to various choices such as the projection of a selected image on the field of the processed image regardless of the nature of the selected image, (the advertising image, flags image or the product image) as well as the possibility of inserting circles fr the for the follow-up of the players on the field by giving the radius of the circle and an off-side line for a given player (in the case of soccer) which will be projected on the field of the processed image and the insertion of a distance arrow projected on the field of the processed image (in the case of a soccer or handball field) by measuring the distance in meters separating the the goal (right or left) from a chosen position and drawing up this arrow projected on the field according to the dimensions in meters of the selected original field.

In addition, this method provides the possibility of writing a statement or the score of the match for the objects projected directly on the plan of the field of the selected image depending on the selected position.

Algorithm of the method of graphics overlay (chroma keying) on field’s plan

0) Start

1) Choose the position (xp, yp) of the integration and projection of the graphic (image to be projected, circle, off-side line, distance arrow, match score) onto the image field which is being processed and compute the coordinates (x’p, y’p) of its equivalent image on the original image of the corresponding field (football, handball, basketball or handball) and multiply them using the homography matrix «H» :

- x’p = (H(1,1) * xp + H(1,2) * yp + H(1,3)) / (H(3,1) * xp + H(3,2) * yp + H(3,3))
- y’p = (H(2,1) * xp + H(2,2) * yp + H(2,3)) / (H(3,1) * xp + H(3,2) * yp + H(3,3))

2) Insert the selected graphic into the field’s original image corresponding to the position (x’p, y’p):

- In the Case of the integration of a selected image to be projected , we have used the summation of the pixels of the two images, choosing a given degree of transparency.

- In the case of the projection of a circle on the field’s plan, we declare a graphic from the field’s original image and draw a circle having a radius computed in pixels according to the following formula:

*Radius size in pixels = ((Original field’s width in pixels * Size of the radius entered in meters) / Original field’s width in meters)*

- In the case of an insertion of a distance arrow projected on the field, we declare a graphic from the original field’s image and draw on this graphic an arrow having its starting point at the position (x’p, y’p) and arrival point at the position *(0, (Image’s height « img_orign ») /2) in the event of the arrow pointing towards the left side of the Goal, otherwise the arrival point shall be at the position (Image Width « img_orign »,* *(Image Height « img_orign ») /2)* Should the arrow be pointing towards the right side of the Goal, we compute the selected position's coordinates in meters x1 and y1 according to the following formulae:

*x1 = ((x'p * Original Field's Width in meters « img_orign »)*

*y1 = ((y'p * original field's height in meters) / original field's Height « img_orign »)*

Then, we compute the distance « D » in meters for the projected image on the field according to the following formula:

$$\boldsymbol{D} = \sqrt{((\boldsymbol{LTm} - \boldsymbol{x'p})^2 + \ (\boldsymbol{HTm}\ /\ \mathbf{2} - \boldsymbol{y'p})^2)}$$

With « *LTm* » being the width of the field in meters and *« HTm »* being the height of the field in meters.

-In the event of an integration of an off-side line, a graphic is declared using the original field's image and an arrow is drawn on this graphic, having its starting point at the position (x'p, 0) and its arrival point at the position (x'p, *Height of the original field's image*).

3) Compute the matrix «H^{-1}» which represents the inverse matrix of the Homography matrix «H», then proceed to the multiplication of each point P' of the position (x',y') of the original field's image obtained through the «H^{-1}» matrix, the inverse of the homography, to have its corresponding point P of the (x, y) position on the image which is being processed so as to eventually secure the projection of the integrated graphic on the field's plan of the processed image according to the formula $P = H^{1}P'$ making sure that the objects which are in the foreground appear using the filtered image obtained through the first method defined for the filtering of the field's background ,i.e. putting only the pixels which do not exist in the filtered image.

4) The processed image may be saved on a disk or the processing may be cancelled.

5) End

IV. Results of the Application

In this section, we will show examples of field images which we have tested using our solution as well as other methods. With regard to the background color filtering, for each type of pitch, we provide three figures., the first two of which (Figures 2 and 3) give an example of an image filtering according to the method of selection of four background colors of the field image of our solution as well as that of the image of the obtained result. However, the third image (Figure 4) accounts for the image tested using the methods of the Emgu CV library: it is a matter of converting the image into a grey level and then apply to the image the predefined function ***«image.InRange (color1, color2) »*** which makes it possible to eliminate two random colors existing in the obtained image. Then, we apply the "Gaussian" filter in order to have the output image [12].

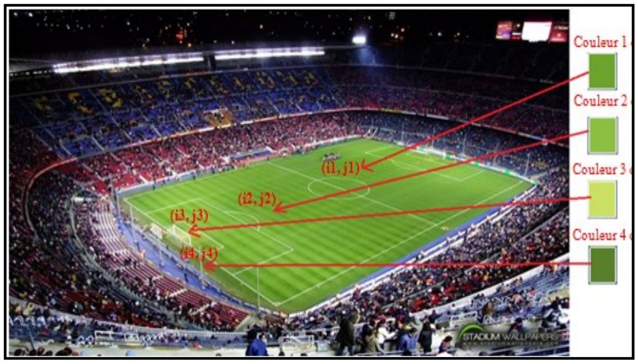


Fig. 2. Example of the selection of the background color chosen for a football pitch filtering

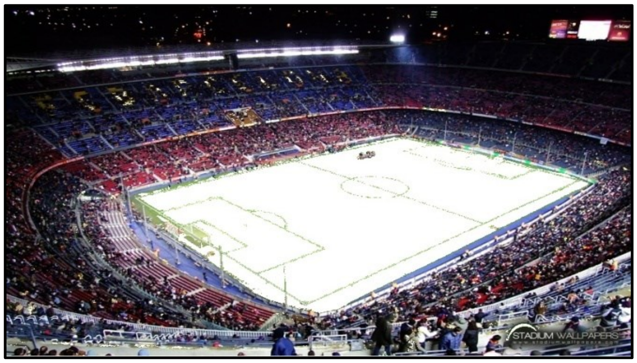


Fig. 3. Outcome of the filtered image of a football pitch according to our method

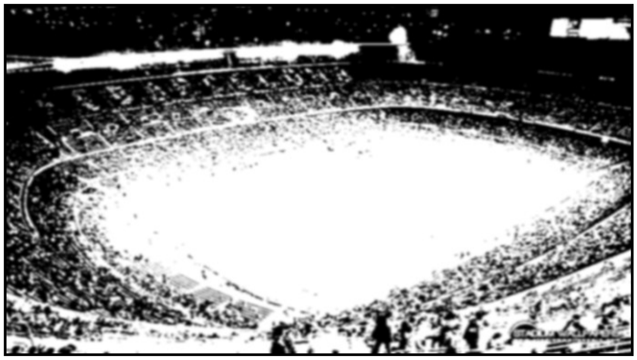


Fig. 4. Outcome of the filtered image of a football pitch according to the use of emgu CV functions

We have made use of the same methods of our solution on tennis courts' images (Figures 5 & 6).

Figure 7 shows the third image which is the image tested according to the Emgu CV library methods.

In the remaining sections of this article, we will be presenting some examples of images which we have tested so as to have the homography matrix proper for each image.

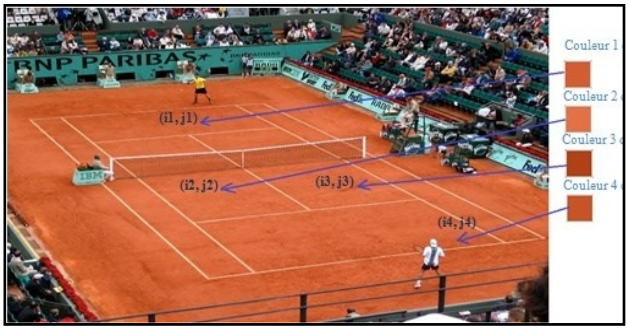


Fig. 5. Example of the selection of the background color chosen for a tennis court filtering

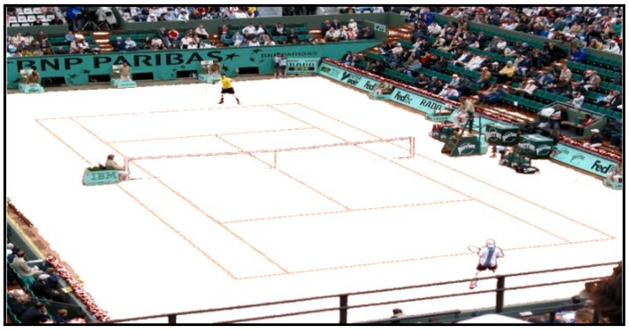


Fig. 6. Outcome of the filtered image of a tennis court according to our method

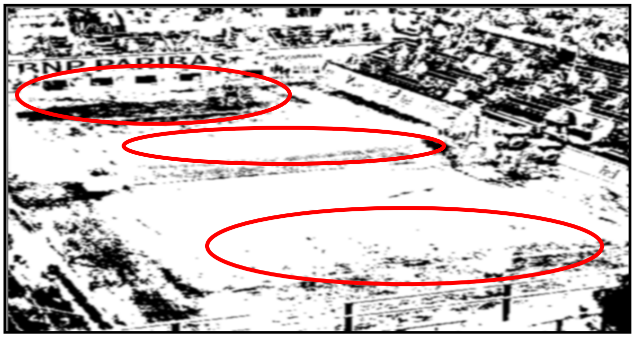


Fig. 7. Outcome of the filtered image of a tennis court tested according to the use of the Emgu CV functions

For every investigated example, we will be specifying four figures, the first two of which represent the list of the coordinates in pixels of selected points and their corresponding and equivalent points in the field's image. Regarding the other two images:

- The first one illustrates the Homography matrix calculated thanks to our programmed algorithm of the used method without resorting to the « Emgu CV » library.
- The second figure illustrates the homography matrix obtained through the use of the predefined «CvInvoke.cvGetPerspectiveTransform (list of the source points, list of the destination points, H)» function which will yield us the Homography matrix «H» by using the « Emgu CV » library [13].

Figures 8, 9, 10 and 11 illustrate an example of a processed football pitch image.

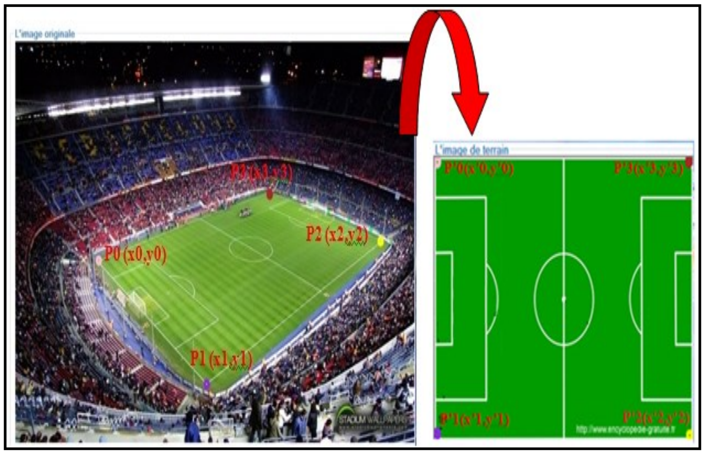
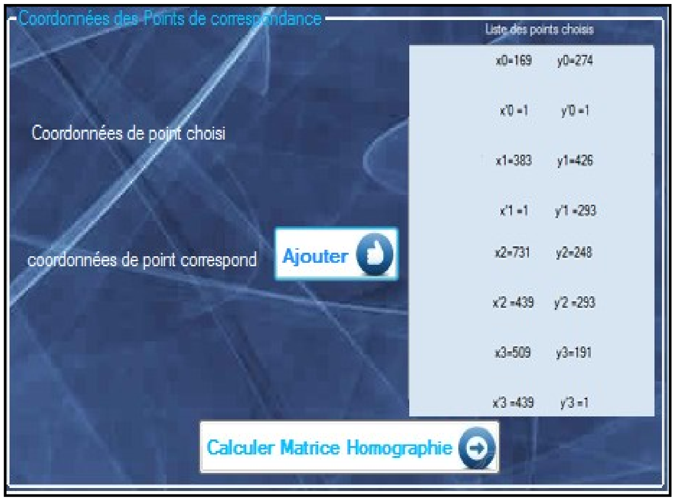


Fig. 8. Example of the selection of the football image points and their corresponding ones on the football pitch



Coordonnées des Points de correspondance

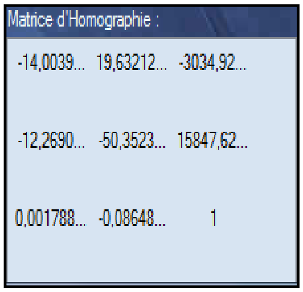
Coordonnées de point choisi

coordonnées de point correspond

Liste des points choisis

x0=169	y0=274
x'0 =1	y'0 =1
x1=383	y1=426
x'1 =1	y'1 =293
x2=731	y2=248
x'2 =439	y'2 =293
x3=509	y3=191
x'3 =439	y'3 =1

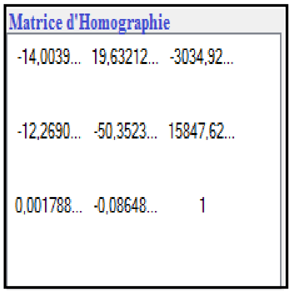
Fig. 9. Display interface of the selected points coordinates list of a football pitch



Matrice d'Homographie :

-14,0039...	19,63212...	-3034,92...
-12,2690...	-50,3523...	15847,62...
0,001788...	-0,08648...	1

Fig. 10. Computed Homography matrix of selected points of a football pitch according to our algorithm



Matrice d'Homographie

-14,0039...	19,63212...	-3034,92...
-12,2690...	-50,3523...	15847,62...
0,001788...	-0,08648...	1

Fig. 11. Computed Homography matrix of selected points of a football pitch according to 'Emgu CV function

Figures 12, 13, 14 and 15 represent an example of a processed tennis court image.

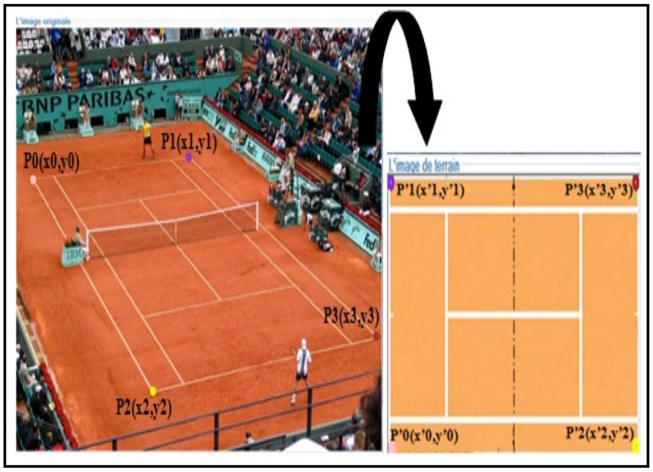
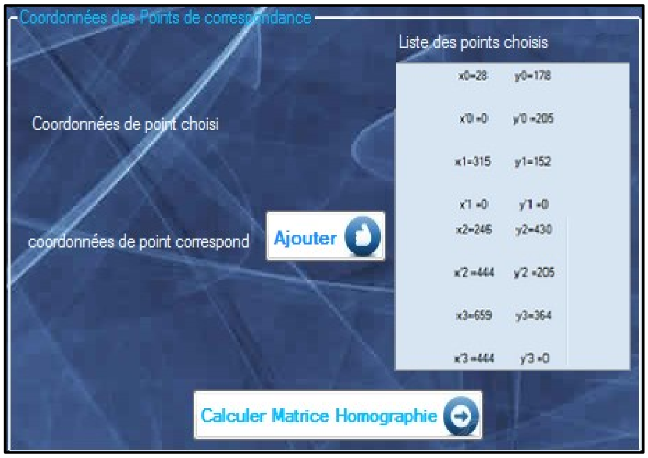


Fig. 12. Example of the selection of the football image points and their corresponding ones on the tennis court


Coordonnées des Points de correspondance
Ajouter
Calculer Matrice Homographie

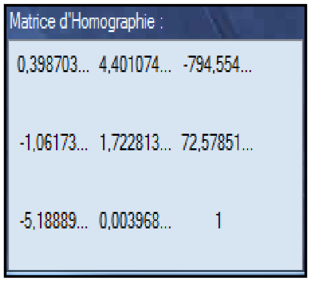
Coordonnées de point choisi

coordonnées de point correspond

Liste des points choisis

x0=28	y0=178
x'0 =0	y'0 =205
x1=315	y1=152
x'1 =0	y'1 =0
x2=246	y2=430
x'2 =444	y'2 =205
x3=659	y3=364
x'3 =444	y'3 =0

Fig. 13. Display interface of the selected points coordinates list of a Tennis court

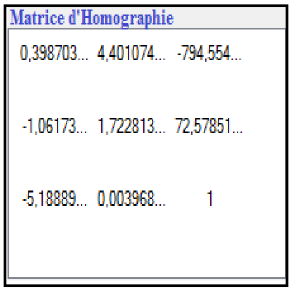

Matrice d'Homographie :

```
0,398703...  4,401074...  -794,554...

-1,06173...  1,722813...  72,57851...

-5,18889...  0,003968...       1
```

Fig. 14. Computed Homography matrix of selected points of a tennis court according to our algorithm


Matrice d'Homographie

```
0,398703...  4,401074...  -794,554...

-1,06173...  1,722813...  72,57851...

-5,18889...  0,003968...       1
```

Fig. 15. Computed Homography matrix of selected points of a tennis court according to the Emgu CV function

Here are examples of images which we have used for the test of the integration of objects on the field plan of each processed image. The aim is to place images of advertising posters (Figures 18 and 19) or flags, insert match scores, draw circles, measure distances of players and insert arrows pointing towards the goal or draw an off-side line on the field according to a chosen position in the case of a football pitch image (Figures 16 and 17).

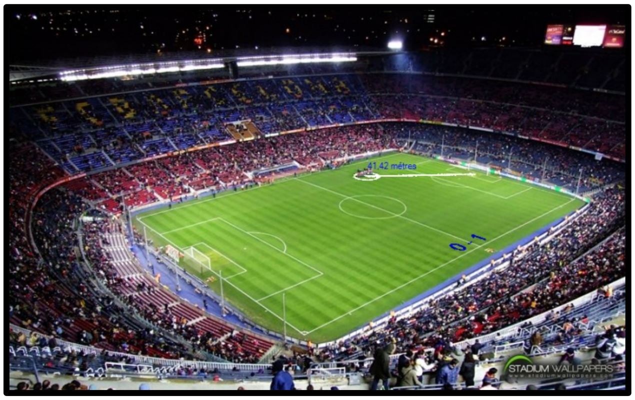


Fig. 16. Insertion of a 9 meter-diameter circle and the distance measurement of the player with the integration of the result projected on the football pitch (example 1)

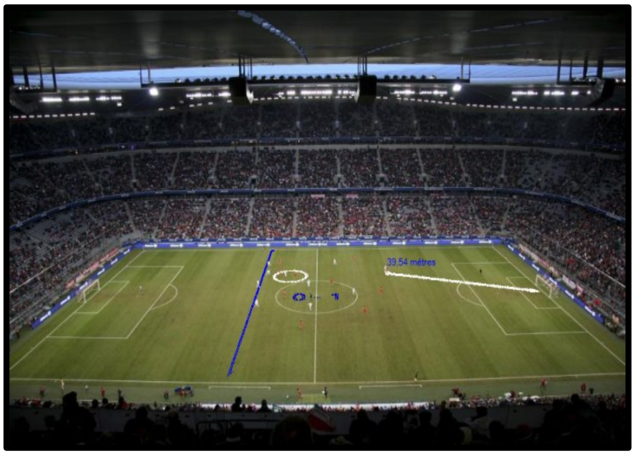


Fig. 17. Insertion of an off-side line and of a 9- meter-diameter circle and distance measurement on the football pitch (example2)

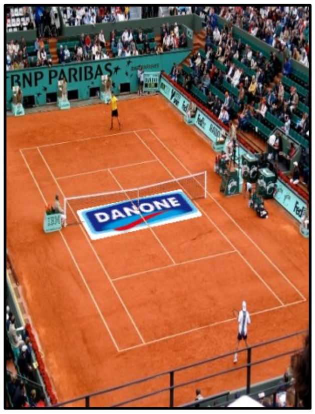


Fig. 18. Integration of an advertising image under the net of a tennis court at a transparency degree = 0%

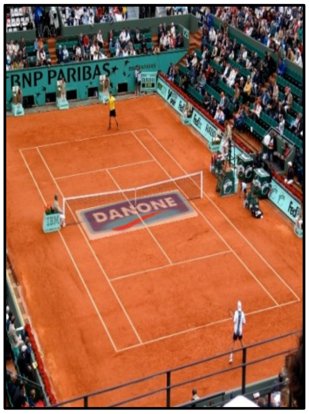


Fig. 19. Integration of an advertising image under the net of a tennis court at a transparency degree = 50%

V. Discussions

According to the earlier results of the background color filtering method, it is worth pointing out that the result of the image filtered according to our method turns out to be better than the result of the filtered field image tested by another method. Thus, it can be said that our background color filtering method is more accurate for the filtering of fields' background color since most of the methods fail to detect the players on the field or the tennis net (See Figure 7)

Moreover, there is a color noise on the field which could be accounted for by an incomplete background color filtering. The implementation time varies according to the efficiency of the algorithm and the features of the used processing machine. Hence, we have carried out a study on the computation and implementation time variation of our background color filtering method algorithm for different images tested with different colors in order to get the result illustrated by the curve in (Figure 20).

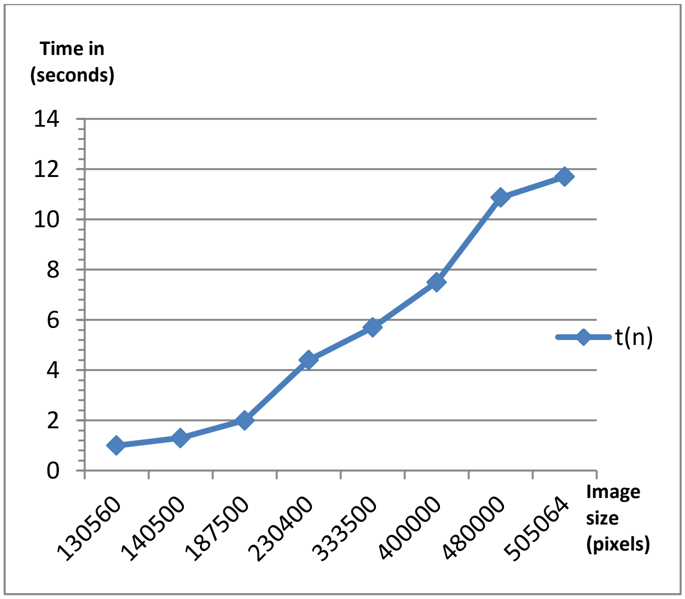


Fig. 20. Implementation time evolution curve of the background color filtering method depending on the image size

According to the results of the examples of the field images tested thanks to the homography matrix computation, we notice that after the plotting of the points and their corresponding ones on the different fields, we find that the result of the Homography matrices obtained and computed by our algorithm (the case of Figures 10 and 14) is equivalent to the result of the Homography matrices obtained using the predefined «Emgu CV» library function (the case of Figures 11 and 15).

Thus, the resulting matrices are the same. We are therefore led to conclude that instead of using the «Emgu CV» library, which suffers some usage problems and shortcomings such as some installation and utilization problems with VB.NET and mainly problems of incompatibility with the operating systems. Thus, we go for the use of our algorithm which makes it possible to give an efficient result at a very fast computation and implementation time (not more than a few milliseconds) whatever the size of the image and the specified points' coordinates.

According to the results of the images tested by the method of the integration of objects on different types of fields, we have noticed that the display of the objects on the field varies according to the homography matrix associated with the processed field's image and we have found that the display takes place in an efficient and more realistic manner, for instance, the results perspective projection on the field of the off-side line (the case of Figures 16 and 17) apart from the integration of advertising images which are projected on the field and under the players as well as under the net of the tennis court (the case of Figures 18 and 19). This makes sure that the integration of what is virtual on what is real of the resulting image is close to the reality. This provides one of the advantages of the augmented reality with the possibility of resizing or the applying of a 90° rotation to manage the orientation of the image projected on the field as well as its size.

Thus, we have tested a method which consists in improving the quality of projection of the result on the field or the integration of lines and circle on the field by applying the Gaussian filter with a 5*5 -size convolution mask which represents the best smoothing filter type of the image for the projection of objects and avoid the noise, if need be, in the image before its projection on the field (Figures 21 and 22).

The Gaussian function is given by:

$$\mathrm{F}\ G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-d\,(x^2+y^2)}{2\sigma^2}}$$

With **σ** being the standard deviation's parameter

It should be remembered that, in general, a Gaussian filter with **σ < 1** is used to reduce the noise and if **σ** is higher than **1, then** the filter will be used, the purpose of which is to construct an image that we could use to make a customized "unsharp mask". Hence, the bigger σ is, the more marked the blur applied to the image will be [14].

Undoubtedly, the major shortcoming of this filter is that the calculation is floating point and not integer at the complexity level [15], [16].

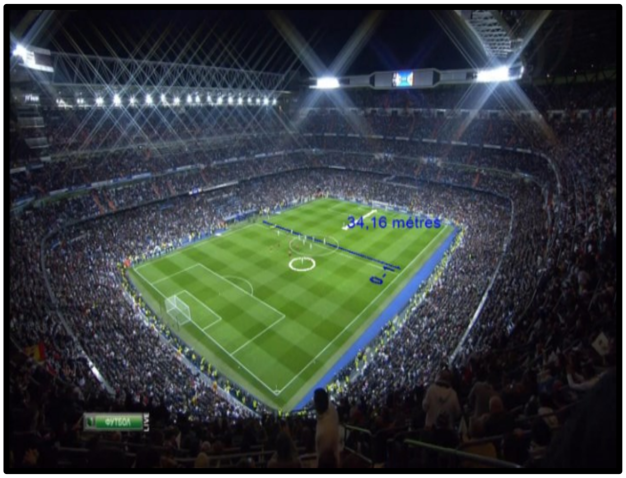


Fig. 21. Result of the objects' integration before the use of the Gaussian filter

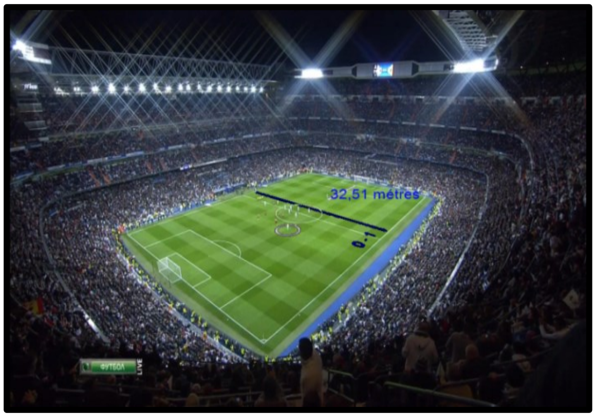


Fig. 22. Result of the objects' integration after the use of the Gaussian filter

So as to secure the efficiency of this method, we have carried out tests on the Handball and Basketball fields' images (Figures 23, 24, 25 and 26).

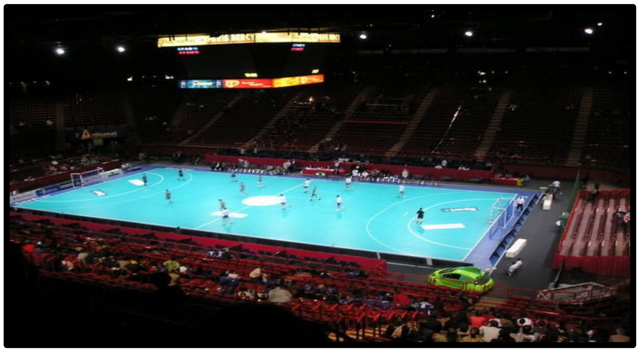


Fig. 23. Original Handball field image to be processed

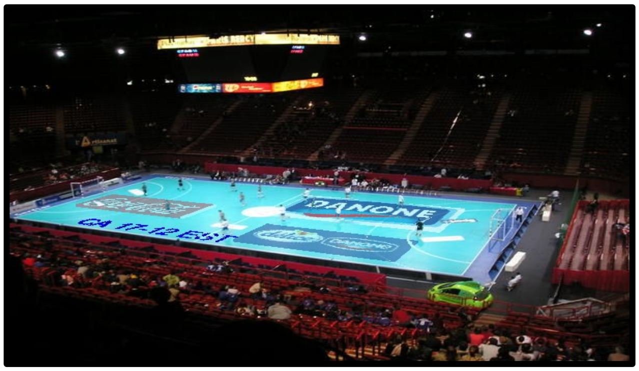


Fig. 24. Insertion of the result and of advertising images under the players on a Handball field at a transparency degree >50%

The performance of the integration computation and of the objects projection on the field time varies according to the size of the image to be processed (in the case of the insertion of a circle, an offside line or a distance measurement) and also depends on the size of the image to be projected but the efficiency of the algorithm is determined thanks to the rapidity of the calculation and implementation time.

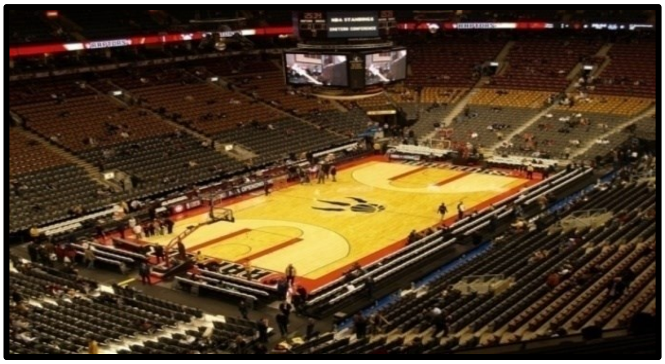


Fig. 25. Original Basketball field image to be processed

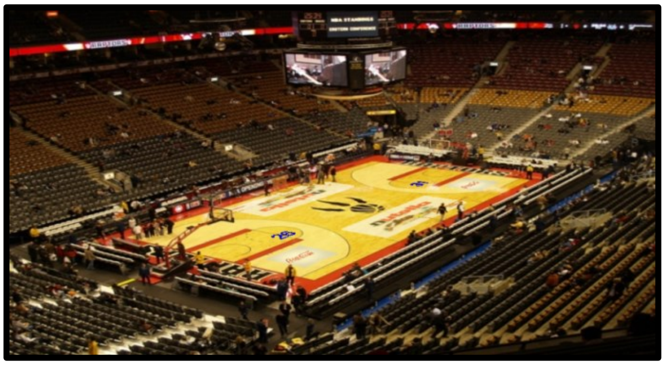


Fig. 26. Integration of advertising images and of scores onto Basketball field

We present the result obtained through a curve which represents the calculation and implementation time in seconds, in relation to the sizes of the images to be projected in pixels with its variation. (Figure 27)

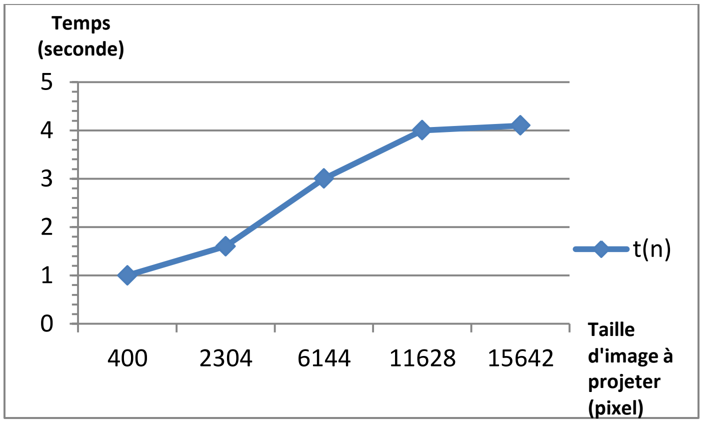


Fig. 27. Curve of the implementation evolution time in relation to the size of the image to be projected on the field.

As to the obtained result which represents the calculation and implementation time in seconds in relation to the sizes of the images processed in pixels with its variation, it is represented by a curve (Figure 28).

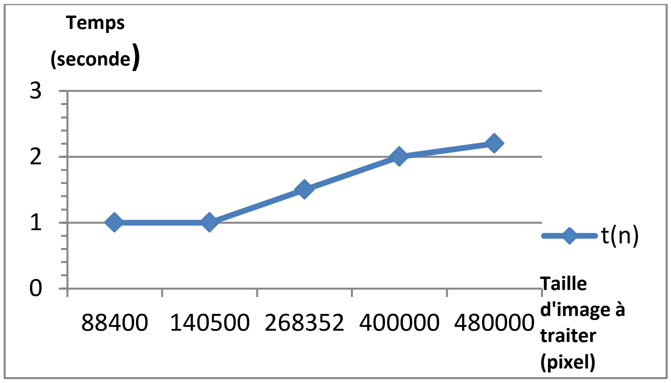


Fig. 28. Curve of the implementation evolution time in relation to the size of the field image to processed.

VI. Conclusions and Outlook

This work demonstrates that our solution has been designed as a system suggested for the augmented reality applied to the telecasting of sports events aimed at analyzing match images where advertising content projected on the real image field plan is inserted so as to eventually create an image superimposing live action and another one calculated from a field model selected in such a way that it integrates virtually with the real image.

It may be concluded that the methods used in our solutions are relevant. To start with, the filtering is more accurate and occurs in the four selected colors efficiently detecting the players and the tennis net on the court to position them in the foreground. Then, the homography matrix calculation is efficient for different types of fields (Football, Tennis, Handball, and Basketball) depending on the models of the fields used. Finally, the overlay of virtual objects on the real output images is the closest one to reality, which is in compliance with the rules of the augmented reality concept.

Our solution, on the other hand, has remained a 2D one, because we have made use of the projection of each object on the selected field's plan at each time. A third dimension Z needs to be included through the use of other Homography matrix calculation techniques and methods so as to be able to inject 3D-images (bottles).

It can also be envisaged to carry out a study considering the possibility of integrating the image distortion correction before processing through using a method which would enable to automatically determine the positions of the field's corners.

The use of this method leading to results deemed to represent the reality should be validated by data generated by real and representative observations so that the integration test occurs in video or in real time.

References

[1] N. Rinous, Word Press. Projet Réalité Augmentée à l'Ensicaen, Etat de l'art.. [accessed 5 May 2015]. available at the following Web address: http://projetar.renous.fr/etat-de-lart/

[2] L. Kim Boyer, S. Sudeep, Perceptual Organization for Artificial Vision Systems, The Kluwer International Series in Engineering and Computer Science, 2000.

[3] T. Ronald, A survey of augmented reality, Presence: Teleoperators and Virtual Environments 6, 4: August 1997, pp. 355-385.

[4] J-M Etienne, Réalité Augmentée et sport : vers des images en 4k, [accessed 5 May 2015]. available at the following Web address: http://www.mediakwest.com/tournage/item/realite-augmentee-et-sport-vers-des-images-en-4k.html

[5] L. Seong-Oh, C.A Sang, J. Hwang, K. Hyoung-Gon, A Vision-Based Mobile Augmented Reality System for Baseball Games, Lecture Notes in Computer Science Vol. 6773, 2011, pp 61-68.

[6] D. Stricker, G. Bleser, From Interactive to Adaptive Augmented Reality, International Symposium on Ubiquitous Virtual Reality (ISUVR), 2012, pp. 18-21.

[7] Rui Yu, Tao Yang, Jiangbin Zheng, Xingong Zhang, Real-Time Camera Pose Estimation Based on Multiple Planar Markers, Fifth International Conference on Image and Graphics, 2009. ICIG '09. 2009, pp. 640-645.

[8] M.I.A. Lourakis, A.A. Argyros , Camera matchmoving in unprepared, unknown environments, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, Vol.2, 2005.

[9] Ji Hoon Choi, Jumyeong Jeok, Seong Yong Lim, Hyun-Cheol Kim, Han-Kyu Lee, Jin Woo Hong, Personalized Data Broadcasting Service based on TV-Anytime metadata, IEEE International Symposium on Consumer Electronics, ISCE 2007, pp. 1-6.

[10] Le Bureau Central de la FIBA Barcelone, Règlement Officiel de Basketball 2014. [accessed 10 May 2015] available at the following Web address: http://www.basketball.qc.ca/images/custom/file/r%C3%A8glement_officiel_du_basketball_2014-09-19.pdf

[11] B. Pierre, USTL. Opérations Géométriques 2D, Cours de Traitement d'Image, [accessed 10 May 2015], available at the following Web address: http://www-lagis.univ-lille1.fr/~bonnet/image/OpGeo.pdf.

[12] M. Bergounioux, HAL. Quelques méthodes de filtrage en Traitement d'Image, [accessed 16 May 2015], available at the following Web address: https://hal.archives-ouvertes.fr/hal-00512280v1/document.

[13] Grenoble Ensimag Kiosk INP. Lissage et filtrage linéaire,. [accessed 6 May 2015], available at the following Web address: http://kevin.polisano.free.fr/Formation/Ensimag2A/TP/ti_sujet_tp2.pdf

[14] Image Convolution, [accessed 16 May 2015], available at the following Web address: http://stephanieluu.com/image-convolution/static1/filtres

[15] EMGU CV Tutorial [en ligne], [accessed 6 May 2015], available at the following Web address: http://www.didehbonyan.com/rz/Portals/0/pdf/Emgu%20CV%20Tutorial%20Skander.pdf

[16] Image Convolution. Filtre moyenneur [en ligne]. [accessed 6 May 2015], available at the following Web address: http://stephanieluu.com/image-convolution/article6/filtre-moyenneur

[17] P. Milgram, F. Kishino. A Taxonomy of Mixed Reality Visual Displays, IEICE Transactions on Information Systems, Vol. E77-D, N°12, 1994.

[18] P. Milgram, H. Takemura, U. Akira, F. Kishino, Augmented Reality: A class of displays on the reality-virtuality continuum. ATR Communication Systems Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun Kyoto 619-02, Japan.1994.

[19] R. Azuma, et al. IEEE. Recent Advances in Augmented Reality, November/December 2001.

[20] L. Maalej, M.N Lakhoua, I. Chakir, T. Battikh and I. Jabri, Planning of a Graphics on TV Project of an Athletics Event, IEEE, CISTEM2014, Tunis 2014.

[21] J-M Cieutat, Quelques applications de la réalité augmentée : Nouveaux modes de traitement de l'information et de la communication. Effets sur la perception, la cognition et l'action. [accessed 6 May 2015] available at the following Web address: https://hal.archives-ouvertes.fr/tel-00802259/document

A Conceptual Framework of Analytical CRM in Big Data Age

Chien-hung Liu
Department of Management Information Management
National Chengchi University
Taipei, Taiwan

***Abstract*—Traditionally, analytical CRM (A-CRM) mainly relies on the use of structured data from a data warehouse where data are extracted, transformed, loaded from operation systems such as ERP, SCM or operational CRM. In recent years of rising big data trend, recognized shifts in E-commerce have taken place from internet-enable commerce (I-commerce), to mobile commerce (M-commerce), and now to ubiquitous commerce (U-commerce).**

As theses paradigm shifts imply that ubiquitous computing improves considerably companies' access to information by allowing them to acquire information at anytime, anywhere. Give this changes on data collection shifts due to ubiquitous computing, however, current A-CRM framework in literature seems not too matched to this change. There is only a handful studies published on CRM in ubiquitous computing environment fitting what big data age requires. Consequently, the objective of this study attempts to propose a conceptual framework of A-CRM. Built by conceptual framework approach, this framework provides valuable directions, definitions and guidelines to practitioners preparing the successful big data marketing in big data age.

Keywords—CRM; Analytical CRM; Big Data; CRM Framework

I. Introduction

Traditionally, analytical CRM (A-CRM) mainly takes advantage of the structured data stored in data warehouse, which collects through operation systems such as ERP, SCM or operational CRM, to create knowledge and insights.

Ubiquitous Computing is already more than a mere technology vision. RFID has already reached a high degree of maturity and is entering more application areas [9]. In recent years, recognized shifts in E-commerce have taken place from I-commerce, to M-commerce, and to ubiquitous commerce (U-commerce) [8]. Kim, Oh and Shin [2] further pointed out that ubiquitous computing improves considerably companies' access to information by allowing them to acquire information at anytime, anywhere.

Give these changes on data collection shifts due to ubiquitous computing, traditional A-CRM frameworks seems not to match up this trend shift. In addition, there are less studies published on CRM in ubiquitous context.

In addition, Ranjan and Bhatnagar [6] believe that traditional operational framework of the customer information has reached its maximum benefit. Likewise, this study believes that the scope of A-CRM should cover different CRM platforms.

Consequently, built on the prior research, the objective of this study aims to propose a conceptual framework of A-CRM with consideration of ubiquitous computing and U-commerce.

This study is organized as follows: (1) introduction ;(2) literature review; (3) research (4) a conceptual framework; and (4) conclusions.

II. Literature Review

A. Ubiquitous computing (UbiComp)

Ubiquitous computing suggests countless very small, wirelessly intercommunicating microprocessors embedded into objects. Equipped with sensors, these computers can record the environment of the objects and provide it with information processing and communication capabilities [9].

Applications of ubiquitous computing are widespread today, for example, retail, personal identification, health care, mobility and transport [9]. In particular, Krumm [5] predicted that ubiquitous advertising will be killer application for the 21st century.

B. U-commerce

Watson, Pitt, Berthon and Zinkhan [13] coined the term "U-commerce" and defined it to as

> *"the use of ubiquitous networks to support personalized and uninterrupted communications and transactions between a firm and its various stakeholders to provide a level of value, above and beyond traditional commerce."*

Wu and Nisa [8] indicated two important traits for U-commerce: ubiquity and universality. "Ubiquity" means that systems can support a rich set of computing and communication capabilities and services while "Universality" means that these systems provide a universal service channel that enables users to stay connected anywhere, anytime, using any devices [8].

In I-commerce stage, internet allows customers to conduct commerce without physical restriction. M-commerce relaxes the independent and mutual constraints of space and time for many commercial activities [13], adding more values then I-commerce. With respect to U-commerce, because it is considered an integration of different channels (from Internet to brick-and-mortar stores), a diversity of ways in which content and services are processed and transmitted [8]. And thus U-commence provides more values then M-commerce and traditional i-commerce.

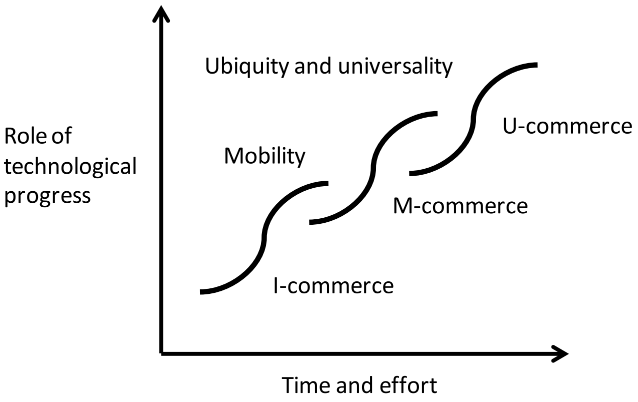


Fig. 1. S CURVES OF EC INNOVATION FROM I-, M-, TO U-COMMERCE [8]

C. *Customer Relationship Management*

In past years, the definition of CRM was not universally agreed among scholars or practitioners. Through years of discussion, definitions for CRM, become more clear. Buttle [4] defines four types of CRM that are mainly used today, depending on the roles that CRM play (Table 1).

TABLE I. TYPE OF CRM [4]

Type of CRM	**Dominant Characteristic**
Strategic	Strategic CRM is a core customer-centric business strategy that aims at winning and keeping profitable customers
Operational	Operational CRM focuses on the automation of customer-facing process such as selling, marketing, and customer service
Analytical	Analytical CRM focus on the intelligent mining of customer-related data for strategic or tactical purposes
Collaborative	Collaborative CRM applied technology across organizational boundaries with a view to optimizing company, partner and customer value

III. RESEARCH METHODOLGY

This study integrated a synthesis of the literature across several academic declines such as Information System (IS), Marketing, Strategy and emerging areas such as big data and ubiquitous commerce to develop a conceptual framework.

This study also use qualitative method interviews with 5 marketing leaders and CRM practitioners from Taiwan to explore and validate how experts think with the proposed framework. These experts that are familiar with CRM were invited to participate. The interviews, ranged from 1 to 2 hours, were conducted during December in 2014. For confidentiality, the name of companies and participants will be presented anonymously.

IV. A CONCEPTUAL FRAMEWORK

This study has developed a conceptual framework for A-CRM with consideration of ubiquitous computing in order to provide with a complete picture of A-CRM.

A. *Loyalty Scheme*

Many scholars tend to view loyalty scheme as a reward program for repeated customers [1][3][4].

This study takes definitions from Liu [1] that a loyalty scheme is a scheme that collects customer information and purchase behaviors, leverages IS technologies to customize rewards for repeated customers so as to develop their long-term profitable loyalty.

The analytical framework starts with customer data collection through loyalty schemes,, regarding its forms, virtual or physical. Loyalty schemes can help business to acquire customers and their customer data, which includes demographics profile, contact information and customer permission of use these data. Therefore, from an analytics point of view, loyalty scheme plays a key role in collecting customer information, which paves the way for CRM and later target marketing.

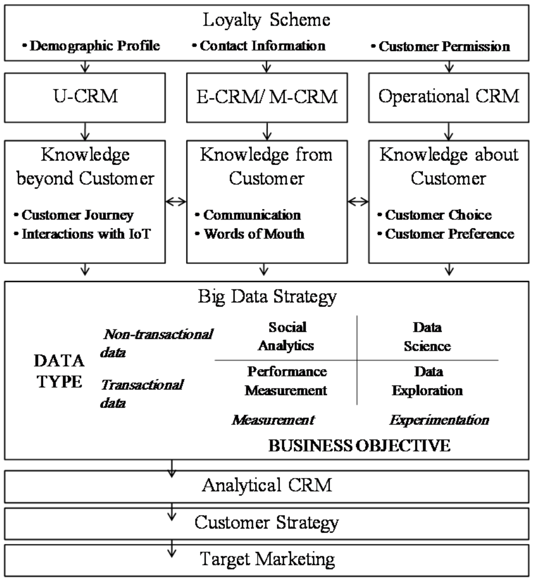


Fig. 2. A CONCEPTUAL CRM FRAMEWORK□

B. *Operational CRM (O-CRM)*

Most of structured data for a conceptual CRM Framework business are collected from Operational CRM (O-CRM) with diverse sources. For example, ERP, SCM, marketing CRM, Sales CRM, call center CRM or etc provide business with knowledge about customers through the transacted data. For example, customer choice and preference can be analyzed through or purchase history associate with customers. In this study, O-CRM, refers to any traditional information systems (IS) that automate customer-facing processes that support selling, marketing and customers services.

C. *Electronic CRM (E-CRM) & Mobile CRM (M-CRM)*

E-CRM is any CRM system that is internet-based. Comparing with traditional O-CRM, E-CRM can supplement O-CRM with non-transactional data in addition to structured data. Data from customers can transactional data through internet or customer communication and interaction data, or even customers' words on the internet.

Mobile CRM is an information system (IS) extending CRM capabilities beyond internet-based platform to mobile devices, which, with its nature of mobility, can collect information from customers, such with location-based, or time-based data and can interact with customers in a real-times manner.

D. *Ubiquitous CRM (U-CRM)*

Wu and Nisa [8] argue that, in recent years, recognized shifts in E-commerce have taken place from I-commerce, to M-commerce, and to ubiquitous commerce (U-commerce) by the development of ubiquitous computing.

Atapattu and Sedera [11] define U-CRM as the use of Ubiquitous Retailing (UR) for their customer relationship management, by retailers specifically. However, ubiquitous technologies are not limited to retail only. For example, Friedewald and Raabe [10] identified other areas of application of ubiquitous computing includes industrial production and management, transport logistics, personal identification and authentication, heath care, mobility and transport.

Therefore, this study terms ubiquitous-CRM (U-CRM) as a CRM system supported by ubiquitous technologies that can sense customer needs and wants, engage customers and used for building long term profitable relationship.

Traditional O-CRM means any information system (IS) that automate the customer-face process that support selling, marketing and customer service. By definitions of Buttle [4], U-CRM, Mobile-CRM, and E-CRM all fill into O-CRM category except the technological platforms based are different.

E. *Big Data Strategy*

Traditional O-CRM provides most of structured data, such as customer transactions, customer choices of products while E-CRM, Mobile CRM and ubiquitous CRM (U-CRM) bring additional new information from customers such as data collected during social communications and interactions, mobility, or customers' word-of-mouth and sentiments, communications between humans, or even devices between devices. This type of data are often heterogeneous and diverse (verity), produced in a real-time fashion (velocity) and big (volume). These three characteristics fit the nature of "big data" described by Doug Laney [9].

Parise, Iyer and Vesset [14] proposed a big data framework for business to shape up strategies to capture and create value. Based on the data type (Y-axis) and business objective (X-axis), a 2x2 matrices is formulated. Big data strategy will provide business with four strategies for value creation from big data, which also can guide the A-CRM.

F. *Analytical CRM (A-CRM)*

Buttle [4] indicates that A-CRM focus on intelligently mining customer-related data for strategical or tactical purposes. Peppers and Rogers [3] thinks A-CRM focuses on the strategic planning needed to build customer value as well as cultural, measurement, and organizational changes required to implement that strategy successfully. Thus, this study believes that A-CRM helps create customer strategy (strategical purpose) that guide target marketing (tactical purpose).

In the U-commerce context, ubiquitous networking is used; The development system is seamlessly integrated with other system; Content design provides needed-based information; , delivery is seamless; services is omnipresent; transactions are multi-disciplined, virtual or physical,; and payment system are diverse [8].

These critical differences imply that business has more opportunity than ever to collect from customers anywhere, anytime, from any devices. This means A-CRM will acquire new data about customer journey, and their communications with other IoT device from U-CRM.

G. *Customer Strategy*

Customer strategies involves examining the existing and potential customer based and identify which forms of segment are most appropriate [12]. This study believes whether a macro, micro, or one-to-tone segmentation approach is appropriate is a decision for a business to make.

H. *Target Marketing*

Targeted marketing is a process where segmentation, targeting, and positioning have to be linked [12]. Peppers and Rogers [3] proposed a "Enterprise Strategy Map" that formulates four strategies based on the level of interacting with individual and level of tailoring products.

At level of granularity of target marketing, one-on-one learning relationship is a cost-efficient target marketing strategy. One-on-one means everyone can be accurately identified, reached with offers or products that are tailored made for each single one of customers. Learning relationship means the data collection through interactions between customers and company as a learning relationship [3]. An increasing number of marketers are looking at personalization to help improve their marketing and expect to bring benefits of one-to-one marketing and customer-relationship management [7].

Interacting	Ability to interact with customer individually	Database marketing	1-1 learning relationship
	Customers addressed only in mass medias	Mass marketing	Niche marketing
		Standard products	Tailored products
		Tailoring	

Fig. 3. Enterprise Strategy Map [3]

V. CONCLUSIONS

Ubiquitous Computing is already more than a mere technology vision. RFID has already reached a high degree of maturity and is entering more application areas [9]. In the ubiquitous age, each individual can be interacted and offered unique tailored product Anywhere, Anytime, with Any device And thus ubiquitous commerce technologies can help customers to get a completely tailored experience with context

This study contributes a conceptual framework for analytical CRM that considers big data happened at a ubiquitous computing environment, which can guide the practitioners or marketers prepare for big data marketing.

This conceptual framework is developed based on combining previous literature, experience as expert opinions. However, this study is not without its limitation. Give a tremendous of own expertise and experience has been devoted to this study, future researcher should consider to include a field-based validation to make this framework more robust and practical as it aims to.

References

[1] C. H. Liu, “A proposal of a typology of loyalty scheme,” A paper accepted by Technology Innovation and Industry Management (TIIM) 2014 Conference, May 28-30, South Korea.

[2] C. Kim, E. Oh and N. Shin, “An empirical investigation of factors affecting ubiquitous computing use of U-business value,” International Journal of Information Management , vol 28, pp. 436-448, 2011.

[3] D. Peppers and M. Rogers, “Managing customer relationship- a strategic framework,” John Wiley & Sons, Inc. Hoboken, NJ, USA, 2011

[4] F. Buttle, “Customer Relationship Management- Concepts and Technologies”, Butterworth-Heinemann, Burlington, MA, USA, 2009

[5] J. Krumm, “Ubiquitous Advertising: The Killer Application for the 21st Century,” IEEE Pervasive Computing, 10(1), 66-73, 2011.

[6] J. Ranjan and V. Bhatnagar, “Role of Knowledge Management and Analytical CRM in Business: Data Mining Based Framework,” The Learning Organization, vol. 18, pp. 131-148, 2011.

[7] J. Versanen and M. Raulas, Building Bridges For Personalization: A process Model for Marketing. Journal of Interactive Marketing, vol. 20, 2006

[8] J. Wu and T. Hisa, “Developing E-business Dynamic Capabilities: An Analysis of E-commerce Innovation From I-, M-, to U-commerce,” Journal of Organizational Computing and Electronic Commerce, vol.18, pp. 95–111, 2008

[9] L. Doug, “3D Data Management- Controlling Data Volume, Velocity, and Varity,” Application Delivery Strategy, META Group, February, 2001

[10] M. Friedewald and O. Raabe, “Ubiquitous computing: An overview of technology impacts,” Telematics and Information, vol 29, pp. 55-65, 2009.

[11] M. Atapattu and D. Sedera, “Ubiquitous Customer Relationship Management: Unforeseen Issues And Benefits,” Pacific Asia Conference on Information Systems (PACIS), 2012

[12] P. Frow and A. Payne, “Customer Relationship Management: A Strategic Perspective,” Journal of Business Market Management, Vol. 3, 7-27, 2009.

[13] R. T. Watson, L. F. Pitt, P. Berthon and G. M. Zinkhan, “U-commerce: Extending the universe of marketing,” Journal of the Academy of Marketing Science, vol. 30, pp. 329–343, 2002.

[14] S. Parise, B. Iyer and D. Vesset, “Four Strategies to Capture and Create Value from Big Data,” Ivey Business Journal, University of Toronto, July/August, 2012.

Design of High Precision Temperature Measurement System based on Labview

Weimin Zhu
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Haima Yang
College of Optical and Electronic Information
University of Shanghai for Science and Technology
Shanghai, China

Jin Liu (Corresponding author)
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Chaochao Yan
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

***Abstract*—Using the LabVIEW software platform, a high precision temperature measuring device is designed based on the principle of the thermocouple. The system uses the STM32 MCU as the main control chip, using AD7076 analog digital converter. The converter has 8 channel, synchronous sampling, and bipolar input. Improving the precision of temperature measurement by cold end compensation, fitting and other measures. The test results show that, the device temperature measurement precision can reach ±0.1 ℃, has the advantages of small size, high precision, and reliable performance, this high precision temperature measurement can be widely used in industrial production.**

Keywords—LabVIEW; AD7076; thermocouple; cold end temperature compensation; Temperature measurement

I. INTRODUCTION

In the industrial production process, temperature is one of the important parameters to measure and control. The conventional temperature measurement method is influenced by the external factors such as emissivity, distance, dust and water vapor,also the measurement error is large. The thermocouple is widely used in the temperature measurement; it has the advantages of simple structure, convenient manufacture, wide measuring range, high precision, small inertia and output signals for transmission and many other advantages. In addition, the thermocouple is a kind of active sensor measurement, and it is not required external power supply and very easy to use, so it is often used to measure the surface temperature of solid and liquid or gas stove. The thermocouple can be used to measure -200 to 1600 ℃ temperature range, and even some thermocouple can measure temperatures above 2000℃.So the thermocouple is one of the most widely used temperature sensor. The temperature measured by thermocouple compensation is a traditional method, effective, the majority of technical staff has accumulated rich experience in the actual measurement.

Virtual instrument measurement technology is becoming more and more important in the field of measurement and control. Virtual instrument can make full use of computer, storage, display and other intelligent functions, through the software, fitting or interpolation correction to solve the cold side and nonlinear compensation. So it is a new topic in the field of temperature measurement that how to combine the thermocouple temperature measurement with the LabVIEW virtual instrument technology. In this paper, the National Instrument Corporation (NI) LabVIEW virtual instrument development platform for thermocouple temperature measurement system with high accuracy and greater application value.

II. THE MEASUREMENT PRINCIPLE AND SYSTEM COMPOSITION

A. The Principle of Temperature Measurement

The temperature measuring system based on thermocouple middle temperature law is the theoretical basis. In the thermocouple cold end potential relationship, the following formula:

$$E_{AB}(t,t_0)=E_{AB}(t,t_1)+E_{AB}(t_1,t_0) \qquad (1)$$

In the formula, the measured temperature is t , the reference temperature is t_0 , the cold end temperature is t_1. In order to facilitate the calculation of the thermocouple indexing table inquiries, we take t_0 as the reference temperature is above 0℃, the formula can be simplified to:

$$E_{AB}(t,0)=E_{AB}(t,t_1)+E_{AB}(t_1,0) \qquad (2)$$

When the cold end temperature $E_{AB}(t,0)$ is 0 ℃ , thermocouple output. When the cold end temperature $E_{AB}(t,t_1)$ is t_1 ℃, thermocouple output. $E_{AB}(t_1,0)$ is the cold end compensation potential. In the formula, $E_{AB}(t,t_1)$ can be directly detected from the thermocouple output. When we get the cold end temperature t_1 , $E_{AB}(t_1,0)$ can be

This work is supported by the National Natural Science Foundation of China under Grant No.61302181, the Natural Science Foundation of Shanghai under Grant No.14ZR1418400 and the Innovation Foundation of Shanghai Education Commission under Grant No13YZ111

calculated by dividing table, thus $E_{AB}(t,0)$ can be calculated. After completing the cold end voltage compensation, measured temperature can be converted by indexing table.

B. Automatic Temperature Measuring System

The structure of the system mainly includes temperature measurement, signal conditioning, data acquisition and AD conversion and PC platform. Among them, the system adopts K type thermocouple to complete temperature measurement, data acquisition and AD conversion using AD7606-F4, MCU control device is used in the design of STMicroelectronics STM32F106VET6, it is a 32 bit microprocessor ARM based on Cortex-M3 kernel.

The structure of component temperature measurement system diagram as shown in figure 1.

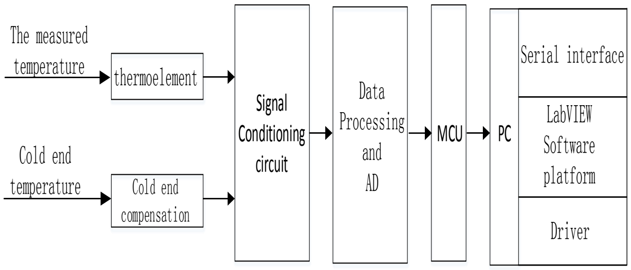


Fig. 1. Structure temperature measurement system

1) Temperature Sensor Selection

The thermocouple as the temperature sensor, which has the advantages of simple structure, easy manufacture, convenient use, high accuracy, in situ measurement and remote measurement, temperature measurement has been widely used in industrial measurement and control system, therefore, this design uses K type thermocouple as temperature measuring element. At the same time, the system uses the Pt100 sensor as the cold end compensating element.

2) The AD Module and Signal Conditioning

Thermocouple sensor signal input differential signal conditioning module, through the amplifier input AD7606, as shown in figure 2.

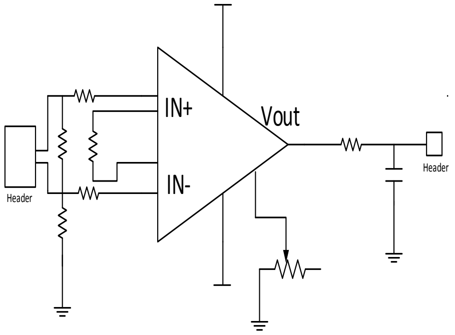


Fig. 2. Signal amplifying circuit

AD7606 synchronous sampling analog-to-digital data acquisition system 16 (DAS), it has respectively 8, 6, 4 acquisition channels. It has on-chip analog input clamp protection, two anti-aliasing filters, track and hold amplifier, 16 bit charge redistribution successive approximation ADC kernel, digital filter, 2.5V reference voltage source and buffer, high speed serial and parallel interface.

In the internal signal conditioning circuit in AD7606, it already contains a low noise, high input impedance signal conditioning circuit, the equivalent input impedance is completely independent of the sampling rate and fixed 1Mohm. At the same time the input terminal integrated with 40dB anti-aliasing filter stack suppression is simplified, the previous design, no longer need the external drive and filter circuit. Therefore, the two signal transformer output can be directly connected to the AD7606, as shown in figure 3.

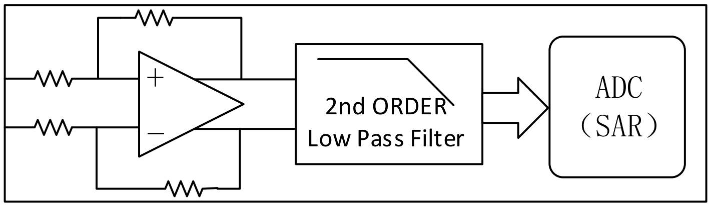
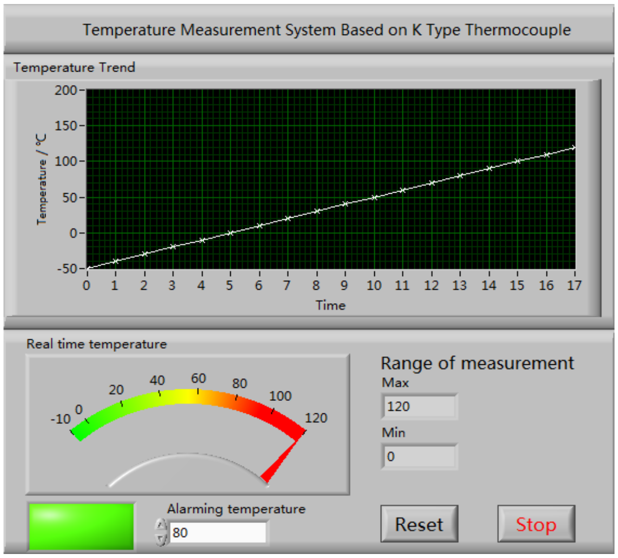


Fig. 3. AD7606 structure design

III. The Software Operation Platform System

The system of virtual instrument technology is based on LabVIEW software. The communication between the PC and the hardware is realized through the serial port. It is the core of software design personalized through the shortcut of the LabVIEW language to achieve the collection, analysis, computing, display and storage. To be able to adapt to the specific requirements of different users, and can continuously adjust the program according to the change of environment or hardware, improvement and optimization of test system, to meet the user's requirements. The use of the software interface as shown in figure 4.


Temperature Measurement System Based on K Type Thermocouple

Range of measurement

Max 120

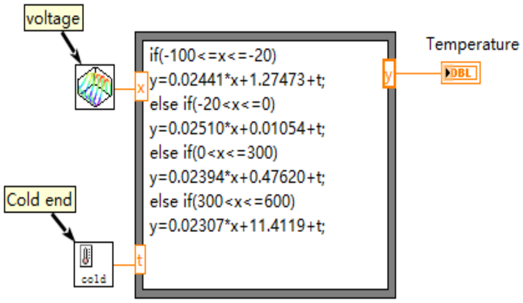
Min 0

Alarming temperature 80

Fig. 4. Program interface

The key issue is the difficulty lies in the design of a full-featured, easy to use, stable performance and friendly interface of the thermocouple temperature testing system. Conversion of voltage signal to the hardware through the verification, determine the voltage is beyond the optimum range, this is because the voltage table temperature is within a certain range, if beyond this range, the accuracy of the thermocouple will not be guaranteed, the measurement is meaningless.

Because of the change of potential temperature K type thermocouple is nonlinear, and the nonlinear lead resistance and other factors, led to the thermocouple output values are deviation from the actual temperature value. Therefore, in order to improve the measurement precision, the data were piecewise linear processing, so as to realize the nonlinear error of the thermocouple calibration. In the temperature range of -100 ℃ ~600 ℃ ~-20 ℃ , divided into -100 -20 ℃ ~0 ℃, 0 ℃ ~300 ℃, 300 ℃ ~600 ℃ by piecewise linear fitting, temperature and thermoelectric potential relationship model, the script and the formula module provided by NI, for the preparation of the formula node module, as shown in figure 5.



```
if(-100<=x<=-20)
y=0.02441*x+1.27473+t;
else if(-20<x<=0)
y=0.02510*x+0.01054+t;
else if(0<x<=300)
y=0.02394*x+0.47620+t;
else if(300<x<=600)
y=0.02307*x+11.4119+t;
```

Fig. 5. Relationship between temperature and thermo emf

IV. Test Results and Analysis

TABLE I. Result and Error of Thermocouple Temperature Measurement

The output voltage value (mV)	The thermostatic bath temperature (℃)	Fitting temperature conversion (℃)	Measurement error (℃)
-2.909	-50	-49.937	0.063
-1.036	0	0.015	0.015
0.984	50	50.047	0.047
3.071	100	99.981	-0.019
5.159	150	150.036	0.036
7.248	200	199.926	-0.074
9.337	250	249.944	-0.056
11.423	300	299.953	-0.047
13.509	350	350.083	0.083
15.67	400	399.937	-0.063
17.842	450	450.049	0.049
20.011	500	500.072	0.072

[a.] The thermostatic bath temperature and analysis

When the cold and hot end voltage data acquisition, floating due to electromagnetic interference or zero drift will cause the voltage, thus showing the temperature values constantly beating, which will lead to a decline in the accuracy of measurement, therefore, paper collected 100 voltage value in 0.1s time, through software programming to solve it the arithmetic mean of the mean, as the sampling results, then the subsequent calculation. This effectively inhibited the beat voltage value, and the accuracy of the measurement results have been better guarantee.

The thermocouple temperature sensor placed in the thermostat, set the temperature of -50 ℃ and the initial test, the thermostat changes with every 50 ℃, the temperature stability, began testing. Experiment results and error as shown in table 1. The results show that the temperature measurement of the temperature measuring device, the absolute error is less than 0.1 ℃, high precision. Have a good practical value in need of high precision temperature measurement.

V. Conclusion

This paper describes the design of a high precision temperature measuring device based on K type thermocouple. Compared with the traditional temperature measuring methods, the device has the advantages of simple circuit structure, high accuracy, good stability. The device can meet the temperature test in the hot test process needs, also has a good prospect in high pressure, high impact and other harsh environments.

References

[1] Q. WANG, Z. G. DONG, and D. W. CHEN. “Implement of High precision Temperature Acquisition System Based on CS5524,”Automation & Instrumentation, 2011, 26(5):58-60.

[2] Z. Q. Sun, J. M. Zhou , and H. J. Sun. “Uncertainty in the Temperature Measurement System Using Thermocouples,” Chinese Journal of Sensors and Actuators, 2007, 20(5):1061-1063.

[3] Z. H. Gao, X. P. Liu, and T. Zhan. “The research of tempration control system based on type-k thermocouple,” Machinery Design & Manufacture, 2011(4):7-9.

[4] Y. ZHANG, S. W. ZHANG. “A Design of High Accurate Temperature Measuring System Based on Platinum Resistance Transducers,” Chinese Journal of Sensors and Actuators, 2010, 23(3):311-314.

[5] M. J. Shi, H. Zhang, and D. Q. He. “Design of High Accurate Platinum Resistance Temperature Measurement System Based on LabVIEW,”2012, 20(4):924-925,938.

[6] J. H. CHENG, B. Qi, and C. B. Qu. “Design and realization of a multichannel temperature measurement system of high-precision and low self-heating,” Transducer and Microsystem Technologies,2014,33(1):56-60.

[7] Y. Qian, Y. M. TANG, and H. C. Yin. “The Design of a Temperature Collection System Based on CPLD,” Chinese Journal of Electron Devices, 2006, 29(2):546-549.

[8] G. Q. Liu, D. H. Tang, and X. W. Li. “Design of a High Accuracy Measuring-temperature System Based on AT89C51,” Chinese Journal of Scientific Instrument, 2005, 26(8):258-262.

Reasoning Method on Knowledge about Functions and Operators

Nhon V. Do
Faculty of Computer Science
University of Information
Technology, VNU-HCM
Ho Chi Minh City, Vietnam

Hien D. Nguyen
Faculty of Computer Science
University of Information
Technology, VNU-HCM
Ho Chi Minh City, Vietnam

Thanh T. Mai
Faculty of Information Technology
Binh Duong University
Binh Duong Province, Vietnam

***Abstract*—In artificial intelligence, there are many methods for knowledge representation. One of the effective models is the Computational Object Knowledge Base model (COKB model), which can be used to represent the total knowledge and to design the knowledge base component of practical intelligent systems. Besides, reasoning methods also play an important role in knowledge base systems. In fact, a popular form of knowledge domain is knowledge about function and operations. These knowledge domains have many practical applications, especially in educational applications, such as Solid Geometry, Analytic Geometry. However, the current methods cannot reason on knowledge about Function and Operators. In this paper, we will present a reasoning method to solve problems on COKB model. These problems are related to knowledge about Functions and Operators. Also, this method has been applied to design some intelligent systems in education. Using this reasoning method, systems can solve problems in some educational knowledge domains automatically with their solutions are step-by-step.**

Keywords—knowledge representation; knowledge based system; intelligent problem solver; automated reasoning

I. Introduction

Intelligent problem solvers (IPS), presented in [1], can consist of components such as theorem proving, inference engines, search engines, learning programs, classification tools, statistical tools, question-answering systems, machine translation systems, knowledge acquisition tools. The IPS system can solve problems in general forms. Users only declare hypothesis and goal of problems base on a simple language but strong enough for specifying problems. The hypothesis can consist of objects, relations between objects or between attributes. It can also contain formulas, determination properties of some attributes or their values. The goal can be to compute an attribute, to determine an object, a relation or a formula. After specifying a problem, users can request the program to solve it automatically or to give instructions that help them to solve it themselves.

An important component in IPS system is knowledge base. This component contains knowledge to solve problems of the system. Knowledge base is designed by knowledge representation methods. Besides the knowledge base, these methods also should be convenient for designing inference engine and interface of the system. Some methods for representation knowledge have been studied. Many classical methods to design knowledge model has been proposed and apply, such as logic, conceptual graph [2], [3], etc. But those results are not efficient for executing and using in the real application.

In real application, a popular component of knowledge domain is knowledge component about operations. Operators between objects of knowledge domain are necessary concepts; they help for representation this knowledge exactly. Besides operators component, knowledge domains can contain relations as function between objects. Example in Knowledge domain about 2D-Analytic Geometry, it contains not only operators between vectors, but also functions about *distance* between two points, between a point and a line, etc. However, the current knowledge models are not effectively for representing knowledge about operators and function. These components cannot be separated but they have to be considered in a knowledge-based system that has many components, this system also includes concepts and relations.

In this paper, Functions and Operators components in model of Computational Objects Knowledge Bases (COKB) will be researched. After that, model of problems will be built and algorithms for reasoning to solve them will be also designed. Moreover, the IPS systems have been constructed; it can solve problems in some educational knowledge domains automatically. The solutions of these systems are naturally, step-by-step, like human's thinking.

The next section will give an overview of the related work. Section III will present a knowledge model about Functions and Operators; this model is constructed based on COKB model. In section IV, problems on this model will be modeled and classified, and algorithms to solve them also will be designed. Based on knowledge model and algorithms in previous sections, the systems of Intelligent Problem Solvers will be constructed in section V. These systems have been applied in some knowledge domains about physics and mathematics. In the last section, we will present conclusion of this study and future work.

II. Related Work

An algebraic structure of elementary knowledge has been presented base on first order logic, as in [4]. This study have solved problem about information equivalent of knowledge. In the other research, in [5], this problem have also solved by constructing automorphic of knowledge-based. However, knowledge domains in these studies are very simple.

Some methods for automated theorem proving has been researched, such as using Groebner bases [6] or algebraic structure based on Temporal logics [7], etc. However, those methods cannot represent knowledge domain precisely. Moreover they cannot implement the way of human's thinking to solve problems.

A modern approach for representing knowledge is ontology. Ontology has been researched and developed in practice application [8, 9, 10]. Ontology COKB has been used in designing some intelligent systems in education (see [11, 12]). In the other study, authors built an ontology for some kinds of relations based on predicate logic and description logic [13], however, this model have drawback to represent complexity relations in human's knowledge.

The study in [14] presents algebraic structure of knowledge about relations; thought that, authors represent a method for knowledge acquisition. However, this model has not yet mentioned structure of concepts and rules.

In fact, a popular form of knowledge domain is knowledge about function and operations. These knowledge domains have many practical applications, especially in educational applications. However, the current methods cannot reason on knowledge about Function and Operators. In [15], authors presented a method for representation this knowledge, but this model is still independence with knowledge about operators and function.

Besides, problems need to be modeled so that we can design algorithms for solving problems automatically. For automatic inference methods, the most important thing is controlling strategy for generating new facts from known facts. Many automatic inference techniques have been studied quite completely in a general level, including: (see [3, 16])

1) Unification routine in methodology for knowledge base representation by first-order logic.

2) Inference by forward chaining

3) Inference by backward chaining

4) Combining forward and backward chaining together with using heuristic rules.

However, the above results are still too general. Some models and techniques are still partial and not good enough for constructing a knowledge system with requirements which are not easy for designer to set up and implement.

In [12, 17], authors presented a reasoning method using Sample Problems. In the processing of solving a practical problem, this method will search for relating problems which were solved before and then apply it to find the solution of the current problem. However, in these models, components about Operators and Functions have not yet researched completely. Some problems about them have not yet been solved, such as: specification of operators and functions, compute an expression between objects, determine value of a function, etc.

UMS Software, in [22], is a program that supports to solve algebraic problems. Some kinds of problem can be solved by this program: solve the equation and basic inequalities, solve the system of equation in two variables, factorization of a Polynomial, simplification of algebraic expression, radical simplification, etc. Besides the solution by textual, UMS Software accompanies the solution with voice comments. But the program is not equipped a knowledge base of algebraic knowledge, it solve problems as frame. The other program in [23] is a website for supporting to solve problems automatically in mathematical knowledge domains. It shows the solutions step-by-step. However, especially in geometry, objects in this program are still simple, and it has limit about number of objects in problems (five objects), so specification about relations between objects is not completely.

III. KNOWLEDGE MODEL ABOUT FUNCTIONS AND OPERATORS

A. Model of Computational Object Knowledge Base

The model of Computational Objects Knowledge Bases (COKB) has been established from the integration of ontology engineering, object-oriented modeling and symbolic computation programming. This way also gives us a method to model problems and to design algorithms. The models are very useful for constructing components and the whole knowledge base of knowledge-based systems in practice.

Definition 2.1: The model of Computational Object Knowledge Base (COKB) consists of six components:

(C, H, R, Ops, Funcs, Rules)

The meanings of the components are as follows:

- **C** is a set of concepts of computational objects.
- **H** is a set of hierarchy relation (IS-A relation) on the concepts.
- **R** is a set of relations between the concepts.
- **Ops** is a set of operators.
- **Funcs** is a set of functions.
- **Rules** is a set of rules.

Each concept in C is a class of Com-objects. The structure Com-Objects can be modeled by **(*Attrs, F, Facts, RulesObj*)**. *Attrs* is a set of attributes, *F* is a set of equations called computation relations, *Facts* is a set of properties or events of objects, and *RulesObj* is a set of deductive rules on facts.

H represents these special relations on *C*. This relation is an ordered relation on the set C, and *H* can be considered as the Hasse diagram for that relation. *R* is a set of other relations on *C*, and in case a relation r is a binary relation it may have properties such as reflexivity, symmetry, etc.

The set *Ops* consists of *operators* on *C*. This component represents a part of knowledge about operations on the objects. Almost knowledge domains have a component consisting of operators. The set *Funcs* consists of functions on Com-Objects. Knowledge about functions is also a popular kind of knowledge in almost knowledge domains in practice, especially fields of natural sciences such as fields of mathematics, fields of physics.

The set *Rules* represents for deductive rules. The rules represent for statements, theorems, principles, formulas, and so forth. Almost rules can be written like the form “if <facts> then <facts>”. In the structure of a deductive rule, <facts> is a set of facts with certain classification. Facts must be classified so that the knowledge component *Rules* can be specified and processed in the inference engine of intelligent systems.

B. *Knowledge about Functions and Operators*

1) Knowledge about Functions

There is wide knowledge and many problems related to the functional component in real knowledge domains, such as knowledge of plane geometry, solid geometry, knowledge of alternating current in physics. For example, intersection between two planes is a line, so Intersection can be a function; the common perpendicular of two skew lines returns a line can be modeled as a function. So that, Funcs component in COKB model is necessary to describe functional knowledge in reality. In [18], authors have researched Funcs component. However, some problems on Functional knowledge have not been mentioned, such as specification of functions, determine value of function.

Structure of a function as followed:

Definitions of functions – form 1:

```
function-def ::= FUNCTION name;
                 ARGUMENT: argument-def+
                 RETURN: return-def;
                 [constraint]
                 [facts]
     ENDFUNCTION;
```

Definitions of functions – form 2:

```
function-def ::= FUNCTION name;
               ARGUMENT: argument-def+
               RETURN: return-def;
               [constraint]
               [variables]
               [statements]
     ENDFUNCTION;
 argument-def  ::= name, <name>: type
 return-def    ::= name : type
statements     ::= statement-def+
statement-def  ::= assign-stmt | if-stmt | for-stmt
asign-stmt     ::= name := expr;
if-stmt        ::= IF logic-expr THEN statements+
                     ENDIF; | IF logic-expr THEN
                     statements+ ELSE statements+
                     ENDIF;
 for-stmt      ::=  FOR name IN [range] DO
                  statements+ ENDFOR;
```

Eg. 2.1: In Knowledge domain about Two dimension analytical geometry, DISTANCE function to compute distance between two point has been defined like this:

```
FUNCTION   <DISTANCE >  :
        ARGUMENT: A::POINT, B::POINT
        RETURN: d::REAL
        constraint:
     d:=$\sqrt{(B.x - A.x)^2 + (B.y - A.y)^2}$
ENFUNCTION
```

2) Knowledge about Operators

In knowledge domains about computation, Ops component is very important to describe them. This component represents operators between objects in knowledge domain. On the basis of operators, knowledge is equipped efficient computational methods for knowledge processing. An operator in this model is a binary mapping: Ci × Ci→ Ci, with Ci is a concept in set C [16]. Structure of an operator as followed:

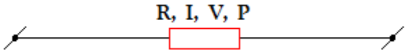
```
operator-def ::= OPERATOR <name>
                  ARGUMENT: argument-def+
                  RETURN: return-def;
                  PROPERTY: prob-type;
                ENDOPERATOR
argument-def ::= name, <name>: type
return-def ::= name : type
prob-type ::= commutative | associative | identity
```

Definition 2.2: Expression of objects is defined like this:

expr ::= object | expr <operator> expr

Rules for determining an operator: Rules-set of Ops-model has been classified two kinds: kind for determining an operator and kind for inference rules of knowledge. The result of an operator is an object of concept in set C. The values of attributes of this object are determined based on rules for determining operator in Rules-set.

Eg. 2.2: In Knowledge domain about Direct Current Electrical Circuit, CIRCUIT concept is modeled by Com-Object, it includes:



Attrs = {R,V,I,P}: set of attributes of CIRCUIT
R,V,I,P: values of resistance, potential difference, current flow, power of circuit.

$$F = \left\{ I = \frac{V}{R}, P = V.I, P = \frac{V^2}{R}, P = I^2R \right\}$$

RulesObj = { }

Operator “ +” performs the series connection between two circuits has been represented like this:

```
OPERATOR  <  +  > :
        ARGUMENT: CIRCUIT, CIRCUIT
        RETURN: CIRCUIT
        PROPERTY: commutative, associative
    ENDOPERATOR
```

Result of this operator has been determined base on rule R1 in section V.A

C. *Basic Techniques*

1) Classify kinds of facts

In the COKB model, there are 12 kinds of facts accepted. These kinds of facts have been proposed from the researching on real requirements and problems in different knowledge domains. The kinds of facts are as follows:

- **Fact of kind 1**: proposition state information about object kind or type of an object.
- **Fact of kind 2**: a proposition states determination of an

object or an attribute of an object.

- **Fact of kind 3**: a proposition states determination of an object or an attribute of an object by a value or a constant expression.
- **Fact of kind 4**: a proposition states equality on objects or attributes of objects.
- **Fact of kind 5**: a proposition states dependence of an object on other objects by a general equation.
- **Fact of kind 6**: a proposition states a relation between objects or attributes of the objects.
- **Fact of kind 7**: a proposition states determination of a function.
- **Fact of kind 8**: a proposition states determination of a function by a value or a constant expression.
- **Fact of kind 9**: a proposition states equality between an object and a function.
- **Fact of kind 10**: a proposition states equality between a function and another function.
- **Fact of kind 11**: a proposition states dependence of a function on other functions or other objects by an equation.
- **Fact of kind 12**: a proposition states a relation between functions.

The last six kinds of facts are related to knowledge about functions, the component **Funcs** in model COKB. The problem below gives some examples for facts related to functions. Problem: Let d be the line with the equation 3x + 4y - 12 = 0. P and Q are intersection points of d and the axes Ox, Oy.

a) Find the central point of PQ

b) Find the projection of O onto the line d.

For each line segment, there exists one and only one point that is the central point of that segment. Therefore, there is a function MIDPOINT(A, B) that outputs the central point M of the line segment AB. Part (a) of the above problem can be represented as to find the point I such that I = MIDPOINT(P,Q), a fact of kind 9. The projection can also be represented by the function PROJECTION(M, d) that outputs the projection point N of point M onto line d. Part (b) of the above problem can be represented as to find the point A such that A = PROJECTION(O,d), which is also a fact of kind 9.

2) Unification of facts

Unification algorithms of facts were designed and used in different applications such as the system that supports studying knowledge and solving analytic geometry problems, the program for studying and solving problems in Plane Geometry. The basic technique for designing deductive algorithms is the unification of facts. Based on the kinds of facts and their structures, there will be criteria for unification proposed. Then it produces algorithms to check the unification of two facts. For instance, when we have two facts *fact1* and *fact2* of kind 1-6, the unification definition of them is as follows: fact1 and fact2 are unified if they satisfy the following conditions

(1) *fact1* and *fact2* have the same kind k, and

(2) *fact1* = *fact2* if k = 1, 2, 6.
[*fact1*[1], {*fact1*[2..nops(*fact1*)]}] =
[*fact2*[1], {*fact2*[2..nops(*fact2*)]}] if k = 6
and the relation in *fact1* is symmetric.
lhs(*fact1*) = lhs(*fact2*) and
compute(rhs(*fact1*)) = compute(rhs(*fact2*)) if k =3.
(lhs(*fact1*) = lhs(*fact2*) and rhs(*fact1*) = rhs(*fact2*)) or
(lhs(*fact1*) = rhs(*fact2*) and rhs(*fact1*) = lhs(*fact2*)) if k = 4.
evalb(simplify(expand(lhs(*fact1*) - rhs(*fact1*) - lhs(*fact2*) + rhs(*fact2*))) = 0)
or evalb(simplify(expand(lhs(*fact1*) - rhs(*fact1*) + lhs(*fact2*) - rhs(*fact2*))) = 0) if k = 5.

IV. Problems and Reasoning Method

A. Model of Problems

Definition 3.1: Let knowledge K = (C, H, R, Ops, Funcs, Rules) as COKB model, model of problem include three sets as followed:

(O, F, G)

In which:

$O = \{O_1, O_2, \ldots, O_n\}$;
$F = \{f_1, f_2, \ldots, f_m\}$;
$G = \{ g_1, g_2, \ldots, g_m \}$

In the above model the set **O** consists of objects, **F** is the set of facts given on the objects, and **G** consists of goals.

The problem will be denoted by **(O, F) → G**

Eg. 3.1: **(**Knowledge domain about Solid geometry)

Problem: Let S.ABC be a triangular pyramid, and I, K be arbitrary points of edges SA, SC respectively such that IK is not parallel to AC. Determine the point of intersection of IK and plane (ABC).

The problem can be modeled by some facts below:

O = {POINT: S, A, B, C, I, K,
Triangle Pyramid: S.ABC}
F = {I belongs to SA, K belongs to SC,
IK is not parallel to AC},
G = {Determine: Point of Intersection of IK and Plane[ABC]}.

Besides that, the goal G in model of problem may be:

1) Determine an object or an attribute (or some attributes) of Object or compute a value of a function relative to objects.

2) Determine an object or an attribute (or some attributes) of Object or compute a value of a function relative to objects and adding some condition.

3) Arguments about relationship between objects according to parameters.

Moreover, each goal of problem on COKB which has presented above is a different kind, so to solve generality problem, we have classified the problem above to three kinds of problems (three class of problems). And each kind of problem will be presented below:

Definition 3.2: Give problem on COKB model has the goal G, with G is determining an object or an attribute (or some attributes) of Object or compute a value of a function relative to objects and *adding some condition*. This problem has the form:

$$(O,F) + (O_{condition}, F_{condition}) \rightarrow G$$

where ($O_{condition}$, $F_{condition}$) is condition (to be added) to determine the goal, that G has to ensure, denote H + $H_{condition}$ → G, in which: H is the hypothesis of the problem, $H_{condition}$ is condition that to G must be guaranteed.

Eg. 3.2: **(**Knowledge domain about 2D-analytic geometry)

Problem: Give two lines d_1:$x + y + 1 = 0,$ d_2: $2x - y - 1 = 0.$ Write a equation of a line (d), known (d) through M, with M point has coordinates (1;-1), and (d1) intersect with (d_1) and (d_2) at A, B with condition is $2\overrightarrow{MA} + \overrightarrow{MB} = \vec{0}$.

We have some facts below:

O = {POINT: M, A, B, LINE: d1,d2,d},
F = {equation of d1: x+y+1=0,
equation of d2: 2x -y -1=0,
M(1;-1),
d through M,
d intersect d1 and d2 at A, B},
$O_{condition}$ ={},
$F_{condition}$ = $\{2\overrightarrow{MA} + \overrightarrow{MB} = \vec{0}\}$,
G = {WRITE:equation of d}

Definition 3.3: Give problem on COKB model has the goal G, which is an argument about relationship between objects according to parameters. This problem has the form:

$$(O, F) + PAR \rightarrow G$$

where, PAR is a set of parameter and goal G is an argument about relationship between objects according to parameters.

Eg. 3.3: **(**Knowledge domain about 2D-analytic geometry)

Problem: Give a circle (C), equations of (C) is $(x - 2)^2 + (y - 3)^2 = 4$, and give A, B are two Point, A has coordinates (4;2), B has coordinates (0,m), with m is parameters of problem. Determine m so that AB cut (C).

O = {POINT:A, B, CIRCLE:C},
F = {equation of (C): $(x - 2)^2 + (y - 3)^2 = 4,$
A(4;2), B(0,m)},
G = {Determine: m so that AB cut (C)}.

B. *Reasoning for problem solving:*

Definition 3.4: Give a problem (O, F) → G on model COKB, and the goal G is to determine (or compute) attributes; M is the set of attributes considered in the problem, A ⊆ M. Denote L be the set of facts in the hypothesis.

- Each f ∈ Rules, each O_i ∈ O, each fun ∈ *Funcs*, each ops ∈ *Ops*, each eq_i ∈ *Equations* we define:

f(A) = A ∪ M_A(f), with M_A(f) is set of facts can be deduced from A by f.

O_i(A) = A ∪ $M_A(O_i)$, with $M_A(O_i)$ is set of facts deduced from A by O_i

fun(A) = A ∪ M_A(fun), with M_A(fun) is set of facts deduced from A by fun.

ops(A) = A ∪ M_A(ops), with M_A(ops) is set of facts deduced from A by ops.

eq_i(A) = A ∪ $M_A(eq_i)$, with $M_A(eq_i)$ is set of facts deduced from A by eq_i.

Each eq_j ∈ *Equations* | ($eq_i \neq eq_j$) :

eq_i(A),eq_j(A) = A ∪ $M_A(eq_i,eq_j)$, with $M_A(eq_i,eq_j)$ is set of facts deduced from A by (eq_i,eq_j).

a) Suppose D = [d_1,d_2, …, d_m] is a list of elements, which d_j ∈ *Rules* or d_j ∈ *O* or d_j ∈ *Ops* or d_j ∈ *Funcs* or d_j ∈ *Equations*, that used to solve problem. Denote:

F_0 = F, F_1 = $d_1(F_0)$, F_2 = $d_2(F_1)$,..,F_m = $d_m(F_m$-1) and D(F) = F_m.

A problem is called *solvable* if there is a list D such that G ⊆ D(L).

The process from state F_i (by applied d_{i+1}) to state F_{i+1} is called a *reasoning step* (i=$\overline{0 \dots m-1}$).

b) Suppose a problem (O, F) → G is *solvable* and called S = [$step_1$, $step_2$,…, $step_m$] is list step of solution of problem. Each $step_i$ is *a step of solution* (i=1..m), it has the form:

(r, kf, nf)

r ∈ D: is rule can be applied to produce new facts or new objects, kf is set of known facts, which can be applied in r, nf is set of new facts, which was produced by r(kf).

When dealing with a practical problem, a convenient way to proceed is considering whether we have met a similar or related problem before or not. If so, then the solution for the problem can be obtained effectively. Or we determine the result of relating problems can be used to solve the practical problem. The related problem like this is called Sample Problem [12].

Definition 3.6: Give S is sample problem in Knowledge K of the form COKB model, consists of two parts:

(Psp, Ssp)

Psp is problem of S, it has form (Op, Fp) → Gp (the model of problem on COKB). Ssp is a list solution-step (by *definition 3.4*) , it has the structure Ssp = [$step_{1,}$ $step_{2,}$ $step_{3,\dots,}$$step_k$]. And we call facts(S) is set of hypothesis of S and goal(S) is set of goal of S, where facts(S) = {Op} ∪ {Fp}, goal(S) = {Gp}.

Problem P on COKB model is classified to three kinds (definition 3.1-3.3). This problem can be solved by using forward reasoning with heuristic rules to select rules or sample problems. There are alogithms to solve kinds of problem as followed:

Algorithm 3.1: Give a problem P1 = (O, F) → G on COKB model as definition 3.1, denote: H → G, *Problem P1 on COKB domain has been found by the following steps*:

- Step 1: Record the element in hypothesis and goal part
- Step 2: Check the goal G, if G is obtained the go to step 7.
- Step 3: Using heuristic rules to select a rule for producing new facts or new objects. If it has found a rule can be applied then go to step 2.
- Step 4: Using heuristic rules to select a sample problem for producing new facts or new objects. If it has found a sample problem can be applied then go to step 2.
- Step 5: Searching for any rule, which can be used to deduce new facts or new objects. If has found a rule can be applied then go to step 2.
- Step 6: Giving conclusion: Solution not found; stop the program.
- Step 7: Reduce the solution found by excluding redundant rules and information in the solution.

In deductive process to find rule can be applied for problem on the COKB, at *algorithm 3.1* we have used some heuristic rules to select rule below:

- Priority use of rules for determining objects.
- Transform objects to objects at a higher level in the hierarchical graph if there are enough facts.
- Use rules for producing new objects that contains elements, which are not in existing objects.
- Use rules for producing new objects that have relationship with existing objects, especially the goal, in necessary situations.
- Try to use deduction rules to get new facts, especially facts that have relationship with the goal.
- If we could not produce new facts or new objects then we should use parameters and equations.There are always new facts (relations or expressions) when we produce new objects.

Algorithm 3.2: The basic procedure for test a sample problem if can be applied.

Giving problem P = (O, F) → G on knowledge K of the form COKB model. Suppose Known_Facts is the set of known fact, which was saved during deduction process to solve problem P, and giving Rules_Sample is the set of sample problem in Knowledge K, have the following structure:

Known_Facts = {kf$_1$, kf$_2$, kf$_3$,…,kf$_n$}, each kf$_i$ is one of twelve kind of facts in K (i=1..n).

Rules_Sample = {sp$_1$, sp$_2$,…,sp$_m$}, each sp$_i$ is a sample problem in knowledge K (i=1…m).

Giving S as definition 3.6 is a sample problem in Knowledge K of the form COKB model (S ∈ Rules_Sample).

S can be applied on problem P has been found by the following procedure:

```
KF ← Known_Facts;
$check \leftarrow false$;
if facts(S) $\subseteq$ KF and goal(S) $\not\subset$ KF then
  $check \leftarrow true$;
end if;
if check=true then
      S is a sample problem can be applied on P to produce news facts; return true;
else
      There is no sample problem found; return false;
end if;
```

$\subseteq$: is symbol of subset, in these cases, it was understood by unification of facts. facts(S) $\subseteq$ KF is return true when facts(S) and KF can be unified.

Algorithm 3.3: With problem P on COKB model, the sample problem can be applied on P has been found by the following procedure: (*heuristics was used for this action*)

```
Sample_found ← $flase$;
SP ← Rules_Sample;
$find$ ← Use Heuristic rules to select S in SP, and S can be applied on P (algorithm 3.2)
if $find = true$ then
      Sample_found ← $true$;
      break;
end if;
if Sample_found then
      S is a sample problem can be applied on P to produce news facts;
else
      S is a sample problem cannot be applied on P to
  produce news facts;
end if;
```

To find a sample problem in set of sample problem to resolve for problem on the COKB, at *algorithm 3.3* we have used some heuristic rules to select a sample problem below:

- Priority use of sample have used with high frequency.
- Priority use of sample for determining objects.
- Priority use of sample for producing new objects that have relationship with existing goal of problem.

Algorithm 3.4: Give a problem P2 on COKB has the form (O,F) + (O$_{condition}$, F$_{condition}$) → G as definition 3.2, denote H + H$_{condition}$ → G. Problem P2 can be resolved by applying algorithm 3.1, and priority using heuristic rules to select sample and rule below:

- Priority use of sample has hypothesis in set of condition of problem.
- Priority use of sample has the goals which have relationship with existing objects in set of condition of problem.

- Priority use of rules has hypothesis in set of condition of problem.
- Priority use of rules for determining objects or fact which have relationship with existing objects in set of condition of problem.

Algorithm 3.5: Give problem P3 on COKB model has the form (O,F) + PAR → G (by *definition 3.3*), denote H + PAR → G. Problem P3 has been found by the following steps*:*

- Step 1: Record the element in hypothesis and goal part
- Step 2: Check the goal G, if G is obtained then find a rule or a sample problem, which has hypothesis part is goal G, and the goal part is subset H then sent this rule (or sample problem) to step 7 and go go step 7.
- Step 3: Using heuristic rules to select a rule for producing new facts or new objects and go to step 2.
- Step 4: Using heuristic rules to select a sample problem for producing new facts or new objects and go to step 2.
- Step 5: Searching for any rules, which can be used to deduce new facts or new objects and go to step 2.
- Step 6: Giving conclusion: Solution not found, and stop program.
- Step 7: Giving R is the result from steps above. in this step we will consider R by argue the goal of R according to parameters. And then output result and stop program.

To find a sample problem in set of sample problem to resolve for problem on the COKB, in this *algorithm 3.5* we have used some heuristic rules to select a rule and sample problem below:

- Priority uses of rules for determine objects which have same object kind in set of goal.
- Priority uses of rules for determine new facts which have same kind of fact in set of goal.
- Priority use of rules for determining objects which have relationship with existing objects in set of goal.
- Priority uses of sample for determine new objects which have same object kind in set of goal.
- Priority uses of sample for determine new objects which have relationship with existing objects in set of goal.

V. Application

The structure of Intelligent Problem Solvers (IPS) system and processing to build them has been presented in [1]. It consists of following components:

- The knowledge base.
- The inference engine.
- The explanation component.
- The working memory.
- The knowledge manager.
- The interface.

Knowledge Bases contain the knowledge for solving some problems in a specific knowledge domain. It must be stored in the computer-readable form so that the inference engine can use it in the procedure of automated deductive reasoning to solve problems stated in general forms. They can contain concepts and objects, relations, operators and functions, facts and rules.

The Inference engine will use the knowledge stored in knowledge bases to solve problems, to search or to answer for the query. It is the "brain" that systems use to reason about the information in the knowledge base for the ultimate purpose of formulating new conclusions. It must identify problems and use suitable deductive strategies to find out right rules and facts for solving the problem. In an IPS, the inference engine also has to produce solutions as human reading, thinking, and writing.

The structure of IPS:

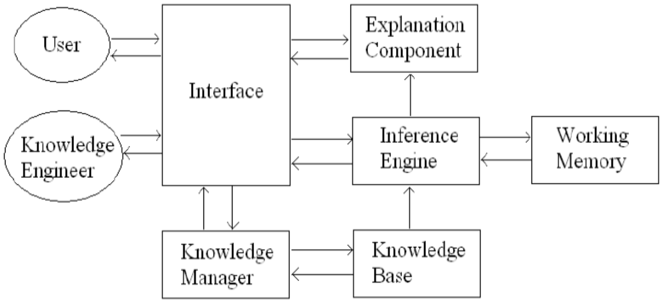


Fig. 1. Structure of a system

The main process for problem solving: From the user, a problem in a form that the user enter is input into the system, and the problem written by specification language is created; then it is translated so that the system receives the working problem in the form of the inference engine, and this is placed in the working memory. After analyzing the problem, the inference engine generates a possible solution for the problem by doing some automated reasoning strategies such as forward chaining reasoning method, backward chaining reasoning method, reasoning with heuristics. Next, the first solution is analyzed and from this the inference engine produces a good solution for the interface component. Based on the good solution found, the answer solution in human-readable form will be created for output to the user.

In this section, we will present some applications in Direct Current (DC) Electrical Circuits, Solid geometry and 2D-analytical geometry. The systems will be designed based on IPS, and knowledge-base component will be modeled and organized by COKB model, the inference engine will be designed based on reasoning algorithms in section III.B.

A. IPS in Direct Current Electrical Circuits

1) Design knowledge-base

Base on knowledge about DC Electrical Circuits has been mentioned in [19], this knowledge domain can be represented by COKB model as followed:

(C, Ops, Rules)

In this model, component H – hicherarchy and component R – relations have been lacked.

Set C of Concepts.

In knowledge domain about DC, Facts component in structure of Com-objects in C is an empty set. So this structure includes three components: **(Attrs, F, RulesObj)**

The set *C* consists of concepts such as “RESISTOR”, “LAMB”, “VariableResistor”, “CIRCUIT”, “CAPACITOR”, “BATTERY”.

<u>Eg. 4.1:</u> Structure of LAMB concept include:

Attrs = {□ R,V,I, P: values of resistance, potential difference, current flow, power of lamb

□ V_{max} , I_{max} , P_{max}: maximum values of potential difference, current flow, power of lamb

□ s: is a number performs three level of light, such as:

s = -1 : the light of lamb is low
s = 0 : the light of lamb is normal
s = 1 : the light of lamb is high}

$$F = \left\{ I = \frac{V}{R}, P = I^2 R, R = \frac{V_{max}^2}{P_{max}}, I_{max} = \frac{P_{max}}{V_{max}} \right\}$$

Rules={if $(I < I_{max})$ or $(P < P_{max})$ or $(V < V_{max})$ then s = -1
if $(I = I_{max})$ or $(P = P_{max})$ or $(V = V_{max})$ then s = 0
if $(I > I_{max})$ or $(P > P_{max})$ or $(V > V_{max})$ then s = 1
}

TABLE I. SET OPS OF OPERATORS BETWEEN CONCEPTS

Operator	Meaning	Arguments	Return	Properties
+	Circuits can be connected by series connection	CIRCUIT x CIRCUIT	CIRCUIT	commutative associative
//	Circuits can be connected by parallel connection	CIRCUIT x CIRCUIT	CIRCUIT	commutative associative
O	Capacitors can be connected by series connection	CAPACICATOR x CAPACICATOR	CAPACI-CATOR	commutative associative
Ξ	Capacitors can be connected by parallel connection	CAPACICATOR x CAPACICATOR	CAPACI-CATOR	commutative associative

In this knowledge domain, RESITOR concept can also be seen as a circuit with one resistor, so operator + and // can be applied on RESITOR concept, these properties and rules of two operators are unchanged.

Set Rules of rules.

Rules in this model are classified of two forms: deductive rules and equation rules.

+ <u>Some deductive rules in Rules-set:</u>

R1: *Rule of series circuits*
{D1, D2, AB: CIRCUIT, AB = D1 + D2}
→ {AB.V = D1.V + D2.V, AB.I = D1.I = D2.I,
AB.R = D1.R + D2.R}

R2: *Rule of parallel circuits*
{D1, D2, AB: CIRCUIT, AB = D1 // D2}
→ {AB.V = D1.V = D2.V, AB.I = D1.I + D2.I,
$\frac{1}{AB.R} = \frac{1}{D1.R} + \frac{1}{D2.R}$ }

R3: *Rule of series capacitors*
{Cap1, Cap2, Cap: CAPACICATOR, Cap = Cap1 o Cap2}
→ {Cap.Q = Cap1.Q = Cap2.Q, Cap.V = Cap1.V + Cap2.V,
$\frac{1}{Cap.C} = \frac{1}{Cap1.C} + \frac{1}{Cap2.C}$ }

R4: *Rule of parallel capacitors*
{Cap1, Cap2, Cap: CAPACICATOR, Cap = Cap1 Ξ Cap2}
→ {Cap.Q = Cap1.Q + Cap2.Q, Cap.V = Cap1.V = Cap2.V,
Cap.C = Cap1.C + Cap2.C}

+ <u>Some equation rules in Rules-set:</u>

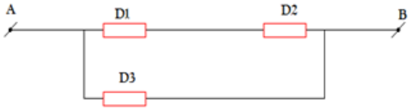
R5: D1 + D2 // D3 = (D1 + D2) // D3
R6: D1 // D2 + D3 = (D1 // D2) + D3
R7: C1 o C2 o C3 = (C1 o C2) o C3
R8: C1 Ξ C2 Ξ C3 = (C1 Ξ C2) Ξ C3

2) Model of problem and algorithm

Model of problem in this knowledge base about DC Electrical Circuits is defined as definition 3.1. Besides that, using algorithms 2.1, the inference engine has been built. This engine simulates the way of human thinking to find solution of practical problem.

3) Illustrating for solving problems

<u>Eg. 4.2:</u> _Three resistors are connected like the figure. Resistor D1 has value 30Ω, Resistor D3 has value 60Ω. The total current is 0.3 A, current though D3 is 0.2A .
What is the current though D1 ?



+ <u>Specification of Problem S:</u>

O:= {D1, D2, D3: RESISTOR ; AB: CIRCUIT}
F:= {AB = D1 + D2 // D3
D1.R=30, D3.R=60, D3.I=0.2, AB.I =0.3}
G:={D1.I}

+ <u>Solution of Program:</u>

STEP 1: D1 + D2 // D3 = (D1 + D2) // D3
by apply rule R5

STEP 2: Let [EF, CIRCUIT], EF = D1 + D2
AB = D1 + D2 // D3
D1 + D2 // D3 = (D1 + D2) // D3
➡AB = EF // D3

STEP 3: AB = EF // D3 → AB.I = EF.I + D3.I
by apply "Rule of Paralle Circuit" R2

STEP 4: EF = D1 + D2 →EF.I = D1.I
by apply "Rule of Series Circuit" R1

STEP 5: From { D3.I = 0.2, AB.I = 0.3, AB.I = EF.I + D3.I} → EF.I = 0.1
by "Solve equation"

STEP 6: From { EF.I = 0.1, EF.I = D1.I} → D1.I = 0.1

B. Intelligent problem solvers in Solid geometry

1) Design knowledge-based

Base on knowledge about Solid Geometry has been mentioned in [20], this knowledge base represented by COKB model consists of five components:

(C, H, R, Funcs, Rules)

The components of the model are listed as followed:

Set C of Concepts.

The set C consists of concepts of Com-objects. There is a variety of concepts in Solid geometry, including "Point", "Segment", "Line", "Angle", "Plane", "Triangle", "Equilateral Triangle", "Isosceles Triangle", "Right Triangle", "Isosceles Right Tri-angle", "Quadrilateral", "Trapezoid", "Right Trapezoid", "Parallelogram", "Rhombus", "Rectangle", "Square", "Triangular Pyramid", "Regular Triangular Pyramid", "Quadrilateral Pyramid" and "Square Pyramid".

Set H of hierarchical relations on the concepts C.

There are hierarchical relationships among the concepts in set C and they can be represented with using Hasse diagrams. For example, the following Hasse diagram demonstrates the hierarchy on the concepts of Quadrilaterals.

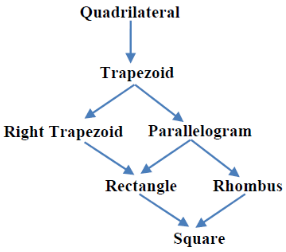


Fig. 2. The hierarchy on the concepts of Triangle

Set R of relations on Com-Objects.

Set R contains various kinds of relations among Com-Objects. The following are some examples of relations: Relation about Perpendicular between a Plane and a Plane, a Line and a Plane, a Line and a Line; Relation about Distinct between a Point and Point, a Line and a Line; Relation about Belong between a Point and a Segment, a Point and a Line, a Line and a Plane; Relation about Altitude between a Segment and a Triangle, a Segment and a Quadrilateral Pyramid.

Set Funcs of functions on Com-Objects.

The set Funcs consists of functions on Com-Objects, such as: Function *Line of Intersection* between two Planes returns a Line. Function *Point of Intersection* between two Lines, or between a Line and a Plane returns a Point. Function *Orthogonal Projection* of a Point onto a Line, or a Point onto a Plane returns a Point. Function *Orthogonal Projection* of a Line onto a Plane returns a Line. Function *Common Perpendicular* of two skew Lines returns a Line. Function *Distance* between a Point and a Line, a Point and a Plane, two Lines, a Line and a Plane, two Planes returns an integer number.

Set Rules of rules.

Rule 1: Three non-collinear points define a plane.
{A, B, C: Point; A, B, C are non-collonear}
$\Rightarrow$ *{Determine a plane (ABC)}*

Rule 2: The intersection point of two diagonals of a parallelogram is the center of the parallelogram.
Parallelogram[ABCD], $O = AC \cap BD \Rightarrow$ *O is center of Parallelogram[ABCD]*

Rule 3: If a point belongs to any lateral edge of a triangular pyramid then the point is not on the base of the triangular pyramid.
SABC is a triangular pyramid; $I \in SA \Rightarrow I \notin plane(ABC)$

Rule 4: If a point belongs to both two distinct lines, then the point is the point of intersection of both of them.
d, d1: Line; H:Point; $H \in d; H \in d1 \Rightarrow H = d \cap d1$

Rule 5: A point belongs to a line and a Plane is the point of intersection of the line and the plane
A: Point; d: Line; P: Plane; $A \in d; A \in P; d \not\subset P \Rightarrow A = d \cap (P)$

Rule 6: If the two distinct points belong to a plane, then the line made by those points lies on the plane.
A, B: Point; P: Plane; $A \in P; B \in P;$ *A, B are distinct* $\Rightarrow AB \subset (P)$

Rule 7: If a point belongs to a line and this line is on a plane, then the point belongs to the plane.
K: Point; d: Line; P: Plane; $K \in d; d \subset (P) \Rightarrow K \in (P)$

Rule 8: If a point is not in a plane then any line containing the point does not lie on the plane. *d: Line; A: Point; P: Plane;* $A \notin P; A \in d \Rightarrow d \not\subset P$

Rule 9: Negative rule.
B, B': Point; B, B' are two distinct points; $B \in d; B = d \cap (P) \Rightarrow B' \notin (P)$

Rule 10: Generated rule a point of intersection of two lines.
K, A, B, C: Point; A, B, C are non-collonear; K is a midpoint of AB; J is a midpoint of AC $\Rightarrow$ *Let* $M = BJ \cap CK$*, M is a midpoint of CK; M is a midpoint of BJ*

2) Model of problem and algorithm

Almost problem in Solid geometry can be modeled by: (O,F) → G, and we can apply algorithm 3.1 for design inference engineer for this application.

3) Illustrating for solving problems

Eg. 4.3: Given a parallelogram ABCD, a point S outside plane(ABCD), a point M between S and A, and a point N between S and B. Let O is intersection point of line AC and BD. Find the intersection of line SO and plane (CMN).

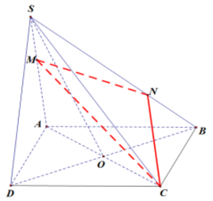


Fig. 3. Problem in eg. 4.3

O = { A, Point], [B, Point], [C, Point], [D, Point], [S, Point], [M, Point], [N, Point], [O, Point], Parallelogram[ABCD]

F = { S ∉ Plane[ABCD], [“Between”, M, Segment[SA]], [“Between”, N, Segment[SB]], O = PointOfIntersection(Line[AC], Line[BD]) }

G = { PointOfIntersection(Line[SO], Plane[CMN]) }

Solution found by the program: Note that, some steps of the following output solution is displayed in the human-readable form instead of using specification language. Therefore, the program helps users get easy to understand the solution.

STEP 1

Parallelogram[ABCD]; O = AC ∩ BD

⇨ O is center of Parallelogram[ABCD]

By rule *“The intersection point of two diagonals of a parallelogram is the center of the parallelogram”*

STEP 2

O is center of Parallelogram[ABCD]

⇨ O= MidPoint(A,C)

By rule *“The properties of the concept Square”*

STEP 3

O= MidPoint(A,C)

⇨ O ∈AC, O is between Segment[AC]

By rule *“The properties of function MidPoint”*

STEP 4

O ∈ AC; AC ⊂ Plane(SAC)

⇨ O ∈ plane(SAC)

By rule *“If a point belongs to a line and this line is on a plane, then the point belongs to the plane”*

STEP 5

M is between Segment[SA]

⇨ M ∈ SA

By rule *“The properties of relation Between”*

STEP 6

M ∈ SA; SA ⊂ Plane (SAC)

⇨ M ∈ Plane(SAC)

By rule *“If a point belongs to a line and this line is on a plane, then the point belongs to the plane”*

STEP 7

O ∈ Plane(SAC); S ∈ Plane(SAC)

⇨ OS ⊂ Plane(SAC)

By rule *“If the two distinct points belong to a plane, then the line made by those points lies on the plane”*

STEP 8

M ∈ Plane(SAC); C ∈ Plane(SAC)

⇨ CM ⊂ Plane(SAC)

By rule *“If the two distinct points belong to a plane, then the line made by those points lies on the plane”*

STEP 9

Triangle[SAC]; O is between Segment[AC]; M is between Segment[SA]

⇨ Let I= OS ∩ CM

By rule *“Generated rule a point of intersection of two lines”*

STEP 10

I = OS ∩ CM

⇨ I ∈ OS, I ∈ CM

By rule *“The properties of function Point of Intersection”*

STEP 11

I ∈ CM; CM ⊂ Plane(CMN)

⇨ I ∈ Plane(CMN)

By rule *“If a point belongs to a line and this line is on a plane, then the point belongs to the plane”*

STEP 12

C ∈ AC; C ∈ Plane(CMN)

⇨ C = AC ∩ Plane(CMN)

By rule *“A point belongs to a line and a Plane is the point of intersection of the line and the plane”*

STEP 13

O ∈ AC; C = AC ∩ Plane(CMN); O, C are distinct

⇨ O ∉ Plane(CMN)

By rule *“Negative rule”*

STEP 14

O ∈ SO; O ∉ Plane(CMN)

⇨ OS ⊄ Plane(CMN)

By rule *“If a point is not in a plane then any line containing the point does not lie on the plane”*

STEP 15

I ∈ OS; I ∈ Plane(CMN); OS ⊄ Plane(CMN)

⇨ I = OS ∩ Plane(CMN)

By rule *“A point belongs to a line and a Plane is the point of intersection of the line and the plane”*

C. IPS in 2D- analytical geometry

1) Design knowledge-base

Base on knowledge about 2D-Analytic Geometry has been mentioned in [21], this knowledge base represented by COKB model:

(C, H, R, Ops, Funcs, Rules)

The components of the model are listed as follows:

Set C of Concepts.

The set C consists of concepts of Com-objects. There is a variety of concepts in Euclid geometry, including “Point”, “Segment”, “Line”, “Angle”, “Triangle”, “Equilateral Triangle”, “Isosceles Triangle”, “Right Triangle”, “Isosceles Right Tri-angle”, “Quadrilateral”, “Trapezoid”, “Right Trapezoid”, “Parallelogram”, “Rhombus”, “Rectangle”, ”Square”, “Circle”, ”Elip”, “Hyperbol” and “Parabol”.

Set H of hierarchical relations on the concepts

There are hierarchical relationships among the concepts in set C and they can be represented with using Hasse diagrams.

For example, the following Hasse diagram demonstrates the hierarchy on the concepts of Triangle.

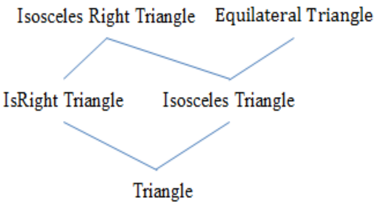


Fig. 4. The hierarchy on the concepts of Triangle

Set R of relations on Com-Objects

Between C-Objects, there are various kinds of relations. We have some relations: Relations of basic level: are relations between basic objects and objects of first level. Relations of first level: are relations between basic objects, objects of first level and objects of second level, or relations between objects of higher levels. Example: Relation Belong between a Point and a Segment, a Point and a Line, a Line and a circle, Segment and Circle, Line and circle…

Ops–set of relations on C-Objects

In knowledge about analytical geometry, we have operators on vectors such as: Operator *create vector* ($\longrightarrow$) from two points it will return a vector. Operator *addition* (+), *subtraction* (-) of vectors it will return a vector. Operator *multiplication* (*) between two vectors, or between a vector with a coefficient.

Set Funcs of functions on Com-Objects

The set Funcs consists of functions on Com-Objects, such as: The set Funcs consists of functions on Com-Objects, such as: Function *distance* between two points it will returns a REAL number. Function *distance* between point and line. Function *symmetry* of a point through a point, or point through line. Function *mid point* of a segment. Function *mid angle* two lines, or two segments, or tow vectors. Function *projection* of a point into a line …etc.

Set Rules of rules

Almost properties, clauses, theorems in geometry can be represented by rules on facts relating to C-Objects. Followings are some particular rules:

{v: vector , u: vector, n: vector, u // v, u $\perp$ n} $\Rightarrow$ { v $\perp$ n}

{u, n: vector, u $\perp$ n} $\Rightarrow$ { u * n = 0},

{AH: segment, ABC:Triangle, AH is altitude of triangle ABC } $\Rightarrow$ {$\frac{1}{AH^2} = \frac{1}{AB^2} + \frac{1}{AC^2}$},

{ABCD: Parallelogram} $\Rightarrow$ {$\overrightarrow{GB} + \overrightarrow{GC} = \overrightarrow{GD}$},

{AD: Segment, ABC: Triangle} $\Rightarrow$ {D is midpoint of BC, BD = DC}, etc

2) Model of problem and algorithms

Model of problems in knowledge domain about 2D-analytic geometry can be represented by models in definition 3.1- 3.3. The inference engine has been also designed base on algorithms correlatively.

3) Illustrating for solving problems

Eg. 4.4: Giving two point B(2;1), C(6;1)

a. Determine A(x,y), with condition, (x>0, y>0) and ABC is equilateral triangle.

b. Determine A' symmetry with A through C.

c. Determine D, with condition $\overrightarrow{AD} - 3\overrightarrow{BD} + 4\overrightarrow{CD} = \overrightarrow{0}$, d. Find M to ABCM is parallelogram. Determine the center of ABCM.

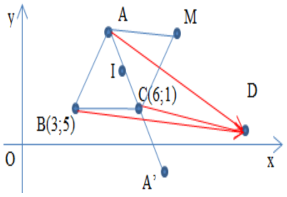


Fig. 5. Problem in Eg 4.4

This problem can be modeled as follows:

O = { [A,POINT],[B,POINT],[C,POINT],
[D,POINT],[M,POINT],[A',POINT],
[I,POINT]}
F = { B=[2,1], C=[6,1]}

$O_{condition}$ = { EQUILATERAL_TRIANGLE[A,B,C],
PARALLELOGRAM[A,B,C,M]
}
$F_{condition}$ =
{VECTOR[AD]-3VECTOR[BD]+4VECTOR[CD]=0,
A'=SYMMETRY(A,C), A.x>0, A.y>0,
["CENTER",I, PARALLELOGRAM[A,B,C,M]]}
G = {Determine: A, A', D, M}

+ Solution of program:

STEP 1

{[EQUILATERALTRIANGLE[A,B,C],
EQUILATERALTRIANGLE]}
→{DISTANCE(A,B)=DISTANCE(A,C),
DISTANCE(A,C)=DISTANCE(B,C),
DISTANCE(A,B)=DISTANCE(B,C)}
By: rule "properties of EQUILATERAL TRIANGLE"

STEP 2

{A=[A.x, A.y], B=[2,1], DISTANCE(A,B)}
→{DISTANCE(A,B)= $\sqrt{(A.x-2)^2 + (A.y-1)^2}$}
By: function "$solving\ of\ function\ distance$" of two points

STEP 3

{ A=[A.x, A.y], C=[2,1], DISTANCE(A,C)}
→{DISTANCE(A,C)= $\sqrt{(A.x-6)^2 + (A.y-1)^2}$}
By: function "$solving\ of\ function\ distance$" of two points

STEP 4

{B=[2,1], C=[6;1], DISTANCE(B,C)}
→{DISTANCE(B,C)=4}
By: function "$solving\ of\ funtion\ distance$" of two points

STEP 5

{DISTANCE(B,C)=4,
DISTANCE(A,B)= $\sqrt{(A.x-2)^2 + (A.y-1)^2}$,
DISTANCE(A,C)= $\sqrt{(A.x-6)^2 + (A.y-1)^2}$,
DISTANCE(A,B)=DISTANCE(A,C),
DISTANCE(A,C)=DISTANCE(B,C),
DISTANCE(A,B)=DISTANCE(B,C)}
→{$(A.x-2)^2+(A.y-1)^2=4$, $(A.x-6)^2+(A.y-1)^2 = 4$}

By: Equation " *auto create equations*" between two equations

STEP 6

{$(A.x-2)^2+(A.y-1)^2=4$, $(A.x-6)^2+(A.y-1)^2 = 4$}
→ {*A*=[4,1+2√3] }
By: Equations "*solve system of equations*" between two equations

STEP 7

{A' = SYMMETRY(A,C), C = [6 , 1], A=[4,1+2√3]}
→{A'=[8,1-2√3]}
By: function "*solving of function symmetry*" of two points

STEP 8

{A = [4 , 1+2$\sqrt{3}$], D = [D.x, D.y]}
→{VECTOR[AD]=[D.x-4, D.y-(1+2√3)]}
By: operators "*create vector*" of two points

STEP 9

{B = [2 , 1], D = [D.x, D.y]}
→{VECTOR[BD]=[D.x-2, D.y-1]}
By: operators "*create vector*" of two points

STEP 10

{C = [6 , 1], D = [D.x, D.y]}
→{VECTOR[CD]=[D.x-6, D.y-1]}
By: operators "*create vector*" of two points

STEP 11

{VECTOR[AD]-3VECTOR[BD]+4VECTOR[CD] = VECTOR[0],
VECTOR[AD]=[D.x-4, D.y-(1+2√3)],
VECTOR[CD]=[D.x-6, D.y-1],
VECTOR[BD]=[D.x-2, D.y-1]}
→{(2D.x-22=0, 2D.y-2- 2√3=0}
By: expression of vector *"compute expression of vector"*

STEP 13

{(2D.x-22=0, 2D.y-2- 2√3=0}
→ {D=[11,1-√3]}
By: System of Equations "*solve system of equations*" between two equations

STEP 14

{[PARALLELOGRAM[A,B,C,M], PARALLELOGRAM]}
→ {VECTOR[AM] = VECTOR[BC]}
By: rules "*properties of parallelogram*"

STEP 15

{A = [4,1+2√(3)], M =[M.x, M.y]}
→ {VECTOR[A,M] = [M.x -4, M.y -1+2√3]}
By: operator "*create vector*" from two points

STEP 16

{B = [4,1], C =[6, 1]}
→ {VECTOR[B,C] = [2,0]}
By: operator "*create vector*" from two points

STEP 17

{VECTOR[AM] = VECTOR[BC]
VECTOR[A,M] = $[M.x - 4, M.y - 1 + 2\sqrt{3}]$
VECTOR[B,C] = [2,0]}
→ {M = [8,1+2√3]}
By: sample problem "two *equal vectors*"

STEP 18

{["CENTER",I,PARALLELOGRAM[A,B,C,M]]}
→ {I=MIDPOINT(A,C)}
By: rule "*properties of parallelogram*"

STEP 19

{I=TRUNGDIEM(A,C),
A = [4,1+2√(3)], C = [6 , 1]}
→{I(5;1+√3)}
By: function "*solving of function midpoint*" of two points.

Eg. 4.5: Giving circle (C) with center I and radius 4(cm) , equation of (C): $(x-2)^2 + (y-3)^2 = 4$, give point: A(5;2), B(0,m), let determine m to AB intersect with O.

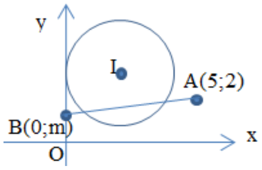


Fig. 6. Problem in Eg.4.5

This problem can be modeled as follows:

O = {[A, POINT], [B, POINT], [I, POINT],
[CIRCLE[I,2],CIRCLE], [LINE[A,B],LINE]}
F = {A=[5;2], B=[0;m]),
C.expr=[$(x-2)^2+(y-3)^2=4$]}
PAR = {m}
G = {
Determine:[m, ["INTERSECTION",LINE[AB],CIRCLE[I,2]]]
}

+ Solution of program:

STEP 1

{ A=[5,2], B=[0,m]}
→{LINE[AB].expr=[(m-2)x+5y-5m=0]}
By: sample problem "*create equation of line*" from two point

STEP 2

{["INTERSECTION",LINE[AB],CIRCLE[C,2]]}
→{DISTANCE(C.I,LINE[AB]) < 2}
By: rule "*relative position of line and circle*" between a line and a circle

STEP 3

{C.expr=[$(x-2)^2+(y-3)^2=4$]}
→{C.I=[2;3]}
By: rule "*properties of circle"*

STEP 4

{DISTANCE(C.I,LINE[AB]),
C. I=[2;3],
LINE[AB].expr=[(m-2)x+5y-5m=0]}
→{DISTANCE(C.I, LINE[A,B])= $\frac{(-3m+11)}{\sqrt{(m-2)^2+25}}$}
By: function "*solving of function distance*" from a point to a line

STEP 5

{DISTANCE(C.I,LINE[AB])< 2,
DISTANCE(C.I, LINE[A,B])= $\frac{(-2m+8)}{\sqrt{(m-2)^2+16}}$}
→ {m>5-2√6}
By: solve of inequality "*solve of inequality according to parmeters*"

STEP 6

Conclude: with {m>5-2√6} then
["INTERSECTION",LINE[AB],CIRCLE[I,2]]]

VI. Conclusion

In previous studies, knowledge about operators and functions have been presented, but they are still discrete and do not research in the model that have many components with concepts and relations between them. In this paper, operators and functions components have been researched in the context of knowledge structure as ontology. Results of this research are: representation, specification, classified kinds of facts relative to operators and functions and processing about unification of facts, model of problems, reasoning methods. With these technologies, scientific basis for designing knowledge base of IPS has been solved.

COKB model with researching about Ops and Funcs component can be used to design and implement intelligent problem solvers. It provides a natural way for representing knowledge domains in practice. By integration of ontology engineering, object-oriented modeling and symbolic computation programming it provides a highly intuitive representation for knowledge. These are the bases for designing the knowledge base of the system. The methods of modeling problems and algorithms for automated problem solving represent a normal way of human thinking. Therefore, systems not only give human readable solutions of problems but also present solutions as the way people write them. COKB and reasoning method on knowledge about Functions and Operators have been applied to construct IPS in education, as: Direct Current Electrical Circuits, Solid Geometry, 2D-Analytic Geometry. The solution given by programs is step-by-step, natural, precise and has reasoning like human.

Our future works are to develop the models and algorithms:

First, different types of COKB model will be researched: reducing of COKB, extension of COKB and integration many types of COKB in real applications.

Second, the rules component of COKB model should have classified by many criteria, such as: deductive rules, equivalent rules, equation rules and reasoning methods base on rules.

Third, researching about relations combines to operators and functions components. In which, it is usefully to research about the coordinate of knowledge about relations, operators and function in knowledge domain that have to using combination these kinds of knowledge.

Acknowledgment

This research in funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2014-26-02.

References

[1] Nhon V. Do. *Intelligent Problem Solvers in Education: Design Method and Applications, Intelligent Systems*, Prof. Vladimir M. Koleshko (Ed.), ISBN: 978-953-51-0054-6, InTech (2012).

[2] Michel Chein & Marie-Laure Mugnier, "Graph-based Knowledge representation: Computational foundations of Conceptual Graphs", Springer-Verlag London Limited, 2009

[3] Chitta Baral, *Knowledge Representation, Reasoning and Declarative Problem Solving*, Cambridge University Press, 2003.

[4] B. Plotkin, T. Plotkin, *An algebraic approach to knowledge base models informational equivalence*, Acta Appl. Math. Vol. **89**, no. 1-3, 109–134 (2005).

[5] Marina Knyazhansky, *Knowledge Bases over algebraic models: Some notes about information equivalence*, Intenational Journal of Knowledge Management, Vol. 8, Issue 1, ISSN: 1548-0666. (2012)

[6] E. Roanes-Lozano, L. M. Laita, A. Hernando, E. Roanes-Macias, *An algebraic approach to rule based expert systems*, Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas (RACSAM), March 2010, Vol. 104, Issue 1, pp. 19-40, Springer (2010)

[7] P. Cordero, G. Gutiérrez, J. Martínez, I.P. de Guzmán , *A New Algebraic Tool for Automatic Theorem Provers*, Annals of Mathematics and Artificial Intelligence, Vol. 42, Issue 4, pp. 369-398 (2004)

[8] Riichiro Mizoguchi, *Advanced course of Ontological Engineering*, New Generation Computing, Vol. 22, Issue 2 (2004), 193-220, Ohmsha Ltd. and Springer-Verlag, 2004.

[9] John F. Sowa, "Knowledge Representation: Logical, Philosophical and Computational Foundations", Brooks/Cole, 2000

[10] Bo Liu, Jianqiang Li, Yu Zhao, "A Query-Specific Reasoning method for Inconsistent and Uncertain Ontology", *2011 International MultiConference of Engineers and Computer Scientists* (IMECS 2011), ISBN: 978-988-18210-3-4, pp. 53-58, Hongkong, March 2011

[11] Nhon V. Do, "*Ontology COKB for designing knowledge-based systems*", Proceeding of 13th International Conference on Intelligent Software Methodologies, Tools, and Techniques (SOMET 2014), pp. 354-373, *New Trends in Software Methodologies, Tools and Techniquies*, Hamido Fujita et al. (eds.), IOS press. (2014)

[12] Nhon V. Do, Hien D. Nguyen, Thanh T. Mai, *Designing an Intelligent Problems Solving System based on Knowledge about Sample Problems*, Proceeding of 5th Asian conference on Intelligent Information and Database Systems (ACIIDS 2013), Kuala Lumpur, Malaysia, March 2013, LNAI 7802, pp. 465-475, Springer

[13] Thomas Bitter, Maureen Donnelly, Computational ontologies of parthood, component hood, and containment, *Proceeding of International Joint Conferences on Artificail Intelligent 2005* (IJCAI 2005), Edinburgh, Scotland, UK.

[14] Jakub M. Tomczak, Jerzy Swiatek, *Knowledge acquisition method for Realational Knowledge Representaion* (2009)

[15] Nhon V. Do, Hien D. Nguyen, *A Knowledge Model about Relations and Application*. Proceeding of 6th IEEE International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM 2012), pp. 707-710, Taipei. 2012

[16] Lakemeyer, G. & Nebel, B., 1994, *"Foundations of Knowledge representation and Reasoning"*. Berlin Heidelberg: Springer-Verlag.

[17] Nhon Do, Hien Nguyen, *A reasoning method on Computation Network and Its applications*, 2011 International MultiConference of Engineers and Computer Scientists, IMECS 2011, ISBN: 978-988-18210-3-4, pp. 137-141, Hongkong , March 2011.

[18] Van Nhon Do, Diem Nguyen, *Intelligent Problem Solving about Functional component of COKB model and Application*. D. Camacho et al. (eds.), *New Trends in Computational Collective Intelligence*, pp. 27-37, Studies in Computational Intelligence 572, Springer (2015)

[19] Vietnam Ministry of Education and Training, *Textbook and workbook of Direct Current Electrical Circuits in middle school*, Publisher of Education (2011).

[20] Vietnam Ministry of Education and Training, *Textbook and workbook of Solid Geometry in high school*, Publisher of Education (2012).

[21] Vietnam Ministry of Education and Training, *Textbook and workbook of 2D-Analytic Geometry in high school*, Publisher of Education (2012).

[22] UMS - Math: http://www.umsolver.com/cgi-bin/club.pl?language=en

[23] Mathway: https://mathway.com/

An Efficient Algorithm to Automated Discovery of Interesting Positive and Negative Association Rules

Ahmed Abdul-Wahab Al-Opahi
Department of Computer Science
Faculty of Computer Sciences and Information Systems,
Thamar University
Thamar, Yemen

Basheer Mohamad Al-Maqaleh
Department of Information Technology
Faculty of Computer Sciences and Information Systems,
Thamar University
Thamar, Yemen

Abstract—**Association Rule mining is very efficient technique for finding strong relation between correlated data. The correlation of data gives meaning full extraction process. For the discovering frequent items and the mining of positive rules, a variety of algorithms are used such as Apriori algorithm and tree based algorithm. But these algorithms do not consider negation occurrence of the attribute in them and also these rules are not in infrequent form. The discovery of infrequent itemsets is far more difficult than their counterparts, that is, frequent itemsets. These problems include infrequent itemsets discovery and generation of interest negative association rules, and their huge number as compared with positive association rules. The interesting discovery of association rules is an important and active area within data mining research. In this paper, an efficient algorithm is proposed for discovering interesting positive and negative association rules from frequent and infrequent items. The experimental results show the usefulness and effectiveness of the proposed algorithm.**

Keywords—Association rule mining; negative rule and positive rules; frequent and infrequent pattern set; apriori algorithm

I. Introduction

Association rules (ARs), a branch of data mining, have been studied successfully and extensively in many application domains including market basket analysis, intrusion detection, diagnosis decisions support, and telecommunications. However, the discovery of associations in an efficient way has been a major focus of the data mining research community [1–2].

Traditionally, the association rule mining algorithms target the extraction of frequent features (itemsets) ie, features boasting high frequency in a transactional database. However, many important itemsets with low support (i.e. infrequent) are ignored by these algorithms.

These infrequent itemsets, despite their low support, can produce potentially important negative association rules (NARs) with high confidences, which are not observable among frequent data items. Therefore, discovery of potential negative association rules is important to build a reliable decision support system. The research in this paper extends discovery of positive as well as negative association rules of the forms A→¬B (or ¬A→B, ¬A→¬B), and so on.

The researchers target three major problems in association rule mining:

a) effectively extracting positive and negative association rules from real-life datasets.

b) extracting negative association rules from the frequent and infrequent itemsets.

c) the extraction of positive association rules from infrequent itemsets.

The rest of this paper is organized as follows. In the second section, related work on association rule mining. In third section, description of interesting positive and negative association rules is presented. The fourth section, the proposed algorithm for discovering interesting positive and negative association rules is described. Experimental results are shown in fifth section. Conclusion and future work are presented in the sixth section.

II. Related Work

A standard association rule is a rule of the form A→ B, where A and B are frequent itemsets in a transaction database and A∩B=Ø. This rule can be interpreted as "if itemset A is true of an instance in a database, so is itesmset B true of the same instance", with a certain level of significance as measured by two indicators, support and confidence. Rule support and confidence are two measures of rule interesting. What if we have a rule such as A→¬B, which says that the presence of A in a transaction implies that B is highly unlikely to be present in the same transaction. Rules of the form A→¬B are called negative rules. Negative rules indicate that the presence of some itemsets will imply the absence of other itemsets in the same transactions [3].

Support-confidence framework for discovering association rules. The validity of an association rule has been based on two measures: the support; the percentage of transactions of the database containing both A and B; and the confidence; the percentage of the transactions in which B occurs relatively only to those transactions in which also A occurs [4].

Investigated the efficient mechanism of identifying positive and negative associations among frequent and infrequent itemsets using state-of the-art data mining technology is presented in [5].

Genetic algorithm GA for mining interesting rules from dataset has proved to generate more accurate results when compared to other formal methods available. The fitness function used in GA evaluates the quality of each rule [6].

Efficacy contemplations for discovering interesting rules from frequent itemsets are suggested in [7-8].

A framework for fuzzy rules that extends the interesting measures for their validation from the crisp to the fuzzy case is presented in [9].

A fuzzy approach for mining association rules using the crisp methodology that involves the absent items is proposed in [10].

Another study introduced to extract interesting association rules from infrequent items by weighting the database "the weight of database must be determined and used the frequent items to discover the infrequent items" [11].

An interesting association rules mining algorithm is proposed to integrate Rule Interestingness measure during the process of mining frequent itemsets, which generates interesting frequent itemsets [12].

Traditional association rules algorithms mostly concentrate on positive association rules. Also, they generate a large number of rules, many of which are redundant and not interesting to the users. The interestingness measures can be used an effective way to filter and then reduce the number of discovered association rules. Based on that, a unified framework is proposed for mining a complete set of interesting positive and negative association rules from both frequent and infrequent itemsets simultaneously

III. Description of Positive and Negative Association Rules

Discovering association rules between items in large databases is a frequent task in knowledge discovery in database KDD. The purpose of this task is to discover hidden relations between items of sale transactions. This later is also known as the market basket database. An example of such a relation might be that 90% of customers that purchase bread and diaper also purchase milk.

TABLE I. Database with 5 Transactions

Transaction	Items
1	Bread,Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Let D be a database of transactions. Each transaction consists of a transaction identifier and a set of items {i1,i2 , ...,in} selected from the universe I of all possible descriptive items. Let D be a database of transactions as shown in Table 1.

The items represents the customer database of sale transactions as a basket data. Each record in this database consists of items bought in a transaction. The problem is how it can be found some interesting (i.e. hidden) relations existing between the items in these transactions or some interesting rules that a manager (a user, a decider or a decision-maker) who owns this database can take some valuable decisions. Some rules derived from this database can {Coke}→{Milk},{Diaper}→{Beer},{Coke,Milk}→{Diaper}.

A positive association rule is an expression of the form: A→B . Each association rule is characterized by means of its support and its confidence defined as follows: Supp (A→B) =Number of transactions containing (AUB) / Total number of transactions. conf (A→B) =supp (A→B) / supp (A). From the above example, rule {Coke}→{Milk} has support 40% and confidence 100%. According to the above measures, the support measure can be considered as the percentage of database transactions for which (AUB) evaluates to be true. The confidence measure is understood to be the conditional probability of the consequent given the antecedent. Association rule mining essentially boils down to discovering all association rules having support and confidence above user-specified thresholds, minsup and minconf, for respectively the support and the confidence of the rules. For example, from the 100% confidence of the rule {Coke},{Diaper}→ {Milk}. It can be concluded that customers that purchase coke and diaper also purchase milk.

In the dataset, it exists other association rule: A→¬B, ¬A→B, ¬A→¬B. The rule A→¬B means the data objects which have itemsets A do not have the itemsets B. The rule ¬A → B means the data objects which do not have itemsets A have the itemsets B. The rule ¬A→¬B means the data objects which do not have itemsets A do not have the itemsets B. These rules can be called negative association rules. For the above example, from the 75% confidence of the rule {Bread} → {¬Coke}. It can be concluded that customers that purchase bread will not also purchase coke. The rule A→B can be called positive association rule. In the existing paper researchers expressed their views on negative association rule in Basket Market database. It is negative association rule which is very useful to the market basket administrator to adjust the business decision making from the customers database. It resolves the lack of past which is only researching positive association rules. This makes the decision makers and access pattern is mined more objective and comprehensive. In order to calculate, the support and confidence for negative association, it can be computed the measures through those of positive rules.

1) $Supp\ (\neg A) = 1\text{-}supp\ (A)$;

2) $Supp\ (A\ U\ \neg B) = supp\ (A)\text{-}supp\ (A\ U\ B)$;

3) $Supp\ (\neg A\ U\ B) = supp\ (B) - supp\ (A\ U\ B)$;

4) $Supp\ (\neg A\ U\ \neg B) = 1\text{-}\ supp\ (A) - supp\ (B) + supp\ (A\ UB)$

5) $Conf\ (A \rightarrow \neg B) = supp\ (A) - supp\ (A\ U\ B)/\ 1\text{-}supp\ (A) =1\text{-}\ conf\ (A \rightarrow B)$;

6) $conf(\neg A \rightarrow B) = supp(\ A\ ) - supp(A\ U\ B)/1\text{-}supp\ (A\ ) = supp\ (B)\text{-}\ supp(\ A\ ) * conf\ (\ A \Rightarrow B\ ) /1 - supp(\ A\ )$

7) $conf(\neg A \rightarrow \neg B) = 1 - supp(A\ ) - supp(B) + supp(A\ U\ B\ ) = conf(\neg A \rightarrow B)/1 - supp\ (A\ )$

8) $Lift\ (A \rightarrow B) = supp\ (A\ U\ B)/\ supp\ (A) * supp\ (B)$

IV. The Proposed Algorithm

The proposed algorithm for automated discovery of interesting positive and negative associations rules consist of two steps:

A. Finding all frequent and infrequent item sets in the database D.

B. Mining interesting association rules (both positive and negative) from the itemsets which we get in the first step.

The interesting measure (lift) has to be greater than one, expressing a positive dependency among the itemsets. The value of lift less than one will express a negative relationship among the itemsets. Figure 1. shows the proposed algorithm.

V. Experiemental Results

The performance of the proposed algorithm on different datasets is demonstrated below and all the codes are implemented under C# language.

A. EXPERIMENT 1

Weather dataset is downloading from the UCI datasets repository. This dataset contains twelf items, fourteen transactions, and seventy words. It helps the researchers in weather forecasts. This datasets applied with varying minsupport and minconfidence values in table 2.We can see that the number of frequent itemsets decreases as we increase the minsupport value. However, a sharp increase in the number of infrequent itemsets can be observed. This can also be visualized in figure 2.

```
1 Generating all the Candidate k-itemsets(Ck)  which its support>0
    1.1  if Ck>= minsupport
         Then   frequent itemsets. Add(Ck)
    1.2  else if Ck<minsupport
         Then Infrequent itemsets. Add(Ck)
2 Generating all interesting association rules from frequent itemsets
    2.1 for each items I in frequent itemsets
    2.2  Generating the rules of the form A==>B
    2.3 If confidence (A==>B)>= minconfidence&& lift(A==>B)>=1
    Then output the rule(A==>B) as frequent positive association rules
3 Generating all the interesting association rules from infrequent itemsets
    3.1 for each items I in Infrequent itemsets
    3.2 Generating the rules of the form A==>B
    else
    Then output the rule(A==>B) as Infrequent positive association rules
    3.3 If confidence (A==>B)>= minconfidence&& lift(A==>B)>=1
    3.4   Generating   the   rules   of   the   form(   A==>~B)
&&(~A==>B)&&(~A==>~B)
    3.5 if confidence(~A==>B)>= mincofidence&& Lift(~A==>B)>=1
    Then output the rule of form(~A==>B) as Infrequent negative rules
    3.6  if confidence(A==>~B)>=mincofidence&& Lift(A==>~B)>=1
    Then output the rule of form(A==>~B) as Infrequent negative rules
    3.7 if confidence(~A==>~B)>= mincofidence&& Lift(~A==>~B)>=1
    Then output the rule of form(~A==>~B) as Infrequent negative rules
```

Fig. 1. The proposed algorithm

TABLE II. Total Generated Frequent and Infrequent Itemsets using Different Support Values

Support	Frequent	Infrequent
0.1	104	136
0.15	42	198
0.20	42	198
0.25	22	218
0.3	11	229

Table 3. gives an account of the experimental results for different values of minimum support and minimum confidence. The liftF value has to be greater than one for a positive relationship between the itemsets; the resulting rule, however, may itself be positive or negative. The total number of positive rules and negative rules generated from both frequent and infrequent itemsets is given. Generates negative association rules of the A→¬B, ¬A→B, ¬A→¬B, which have greater confidence than the user defined threshold and lift greater than one, are extracted as negative association rules figure 3.

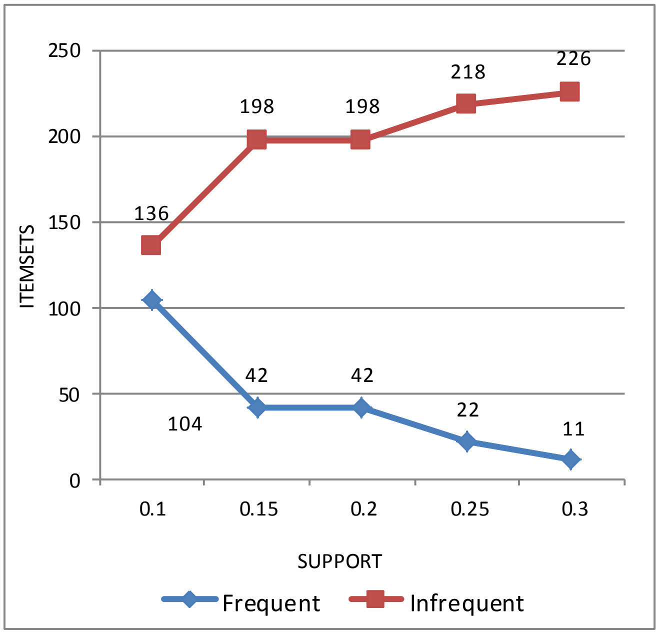


Fig. 2. Frequent and infrequent itemsets generated with varying minimum support values

TABLE III. Intersiting Positive And Negative Association Rules Using Varying Support And Confidence Values With LIFT>1

Supp.	Conf.	PARs From Freq.	PARs From Infrq.	NARs From Freq.	NARs From Infrq.
0.1	0.5	166	351	55	137
0.1	0.7	54	193	15	39
0.15	0.5	35	468	8	134
0.15	0.7	11	236	3	51
0.25	0.5	12	491	0	105

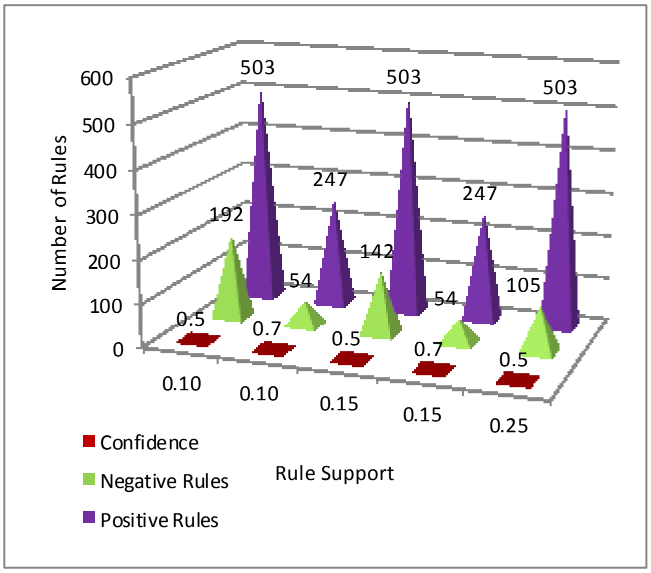


Fig. 3. Intersiting positive and negative association rules generated with varying minimum supports and confidence values

B. EXPERIEMENT 2

The Groceries dataset contains one month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions , the items are aggregated to 169 categories and the total number of words 43367.The frequent and infrequent itemset generation using Apriori algorithm takes only an extra time as compared to the traditional frequent itemset finding using Apriorialgorithm.This is because each item’s support is calculated for checking against the threshold support value to be classified as frequent and infrequent; therefore, we get the infrequent items in the same pass as we get frequent items. The proposed algorithm implemented for Groceries dataset to mine positive and negative from frequent and infrequent items with different parameters (minsupport,minconfidence,3 items length). Table 4. shows that the number of frequent itemsets decreases as it increase the minsupport value. However, a sharp increase in the number of infrequent itemsets can be observed. This can also be visualized in figure 4.

The total number of positive rules and negative rules generated from both frequent and infrequent itemsets which is given in Table 5. Generates negative association rules of the form A→¬B, ¬A→B, ¬A→¬B, which have greater confidence than the user defined threshold and lift greater than one, are extracted as negative association rules figure 5.

TABLE IV. TOTAL GENERATED FREQUENT AND INFREQUENT ITEMSETS USING DIFFERENT SUPPORT VALUES

Support	**Frequent**	**Infrequent**
0.00015	70858	78369
0.00025	44260	104966
0.0005	24172	125025
0.001	9969	139248
0.002	3812	145395

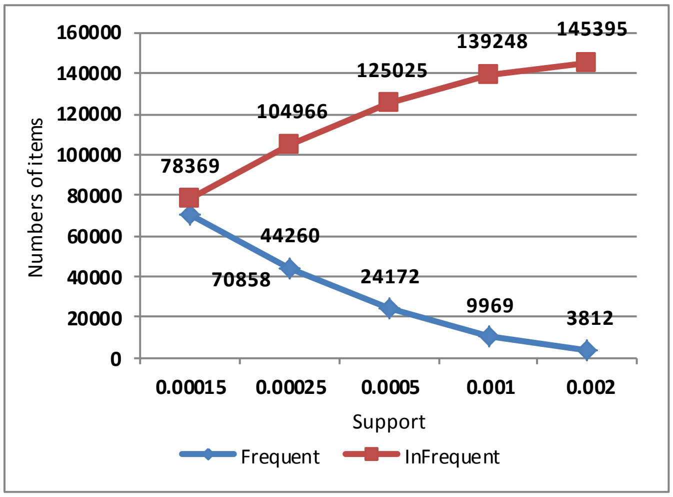


Fig. 4. Frequent and infrequent itemsets generated with varying minimum support

TABLE V. INTERSITING POSITIVE AND NEGATIVE ASSOCIATION RULES USING VARYING SUPPORT AND CONFIDENCE VALUES WITH LIFT>1

Supp.	**Conf.**	**PARs From Freq.**	**PARs From Infrq.**	**NARs From Freq.**	**NARs From Infrq.**
0.0005	50	3772	57496	2451	19190
0.001	50	1472	59835	881	19178
0.001	60	493	31514	881	19178
0.002	50	582	60725	330	15787
0.002	60	138	31869	330	15787

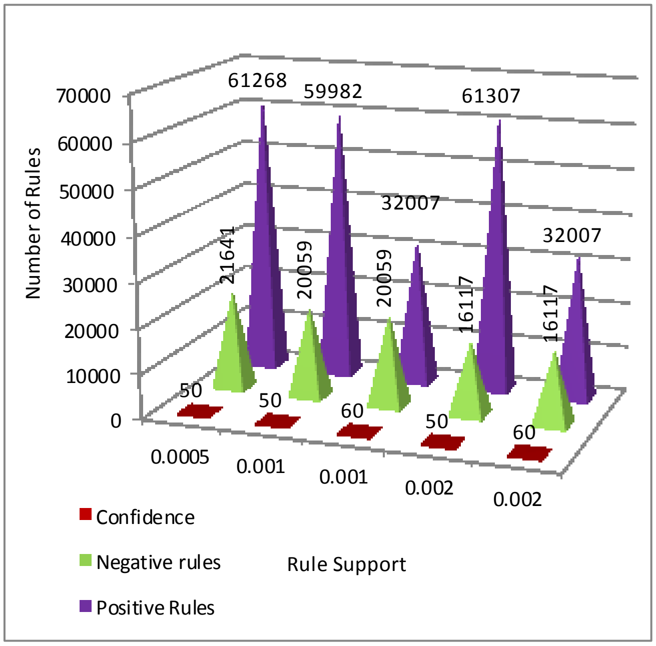


Fig. 5. Intersiting positive and negative association rules generated with varying minimum supports and confidence values

VI. Conclusion and Future Work

In this paper, a new algorithm to generate interesting positive and negative association rules from frequent and infrequent itemsets is proposed. Whereas, traditional association rules mining algorithms have focused on frequent items to generate positive association rules. The proposed algorithm integrates lift as interestingness measure during the process of mining rules. The experimental results have demonstrated that the proposed algorithm is efficient and promising. In future work the researchers will present improved algorithm by using different interestingness measures for mining association rules.

References

[1] F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, "An Improved Algorithm for Mining Association Rules in Large Databases," Vol. 1, No. 7, 311-316, 2011

[2] H. H. O. Nasereddin, "Stream data mining," International Journal of Web Applications, vol. 1, no. 4, pp. 183–190, 2009.

[3] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of bothpositive and negative association rules," ACM Transactions onInformation Systems, vol. 22, no. 3, pp. 381–405, 2004.

[4] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD Record, vol. 22, no. 1, pp. 207–216, 1993.

[5] S. Mahmood, M. Shahbaz, and Z. Rehman, "Extraction ofpositive and negative association rules from text: a temporal approach," Pakistan Journal of Science, vol. 65,pp. 407–413, 2013.

[6] G. pKaur ,and S. Aggarwal, "A Survey of Genetic Algorithm for Association Rule Mining," International Journal of Computer Applications, vol.67, no.20, pp. 19-22, 2014.

[7] T. Slimani, A. Lazzez,"Efficient Analysis of Pattern and Association Rule Mining Approaches," International Journal of Information Technology and Computer Science, vol.6, no.3, pp.70-81, 2014.

[8] E. Duneja,A. K. Sachan,"A Survey on Frequent Itemset Mining with Association Rules,"International Journal of Computer Applications, vol.64, no.23, pp.7105-97202014.

[9] M. Delgado, M. D. Ruiz, D. S´anchez, and J. M. Serrano, "A formal model for mining fuzzy rules using the RL representation theory," Information Sciences, vol. 181, no. 23, pp. 5194–5213, 2011.

[10] M. Delgado, M. D. Ruiz, D. Sanchez, and J. M. Serrano, "A fuzzy rule mining approach involving absent items," in Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology, pp. 275–282, Atlantis Press, 2011.

[11] L. Cagliero and P. Garza, "Infrequent Weighted Itemset Mining Using Frequent Pattern Growth," IEEE Transactions on know ledge and data engineering, vol.26, no.4, 2014.

[12] B. M. Al-Maqaleh and S. S. Ghalib,"Pushing Rule Interestingness Measure in Association Rules Mining," TUJNAS,Accepted, 2015.

Technical Issues and Challenges in Building Human Body Sensor Networks

Meghna Garg,Student
Computer Science
Chitkara University
Baddi(H.P), India

Manik Gupta,Professor
Computer Science
Chitkara University
Baddi(H.P), India

***Abstract*—In this research work, an exploration is done for identification of critical technical issues, problems, challenges in area of wireless body network sensors, which are continuous emerging as integral part of health monitoring systems. All this is possible due to the concept of 'Internet of Things' [5], in which day to day consumer devices, equipment are connected onto the network enabling information gathering and management of many vital signals. The first section gives introduction to wireless network developments in recent context, and then it is followed by discussion on existing work done in context of physical, Media access control and other aspects like energy consumption and security of such systems. The tabular summary on the gaps, limitations of Wireless body area network is done and based on this work; future directions are also suggested. Care has been taken to solicit high impact general paper for conducting this systematic study**.

Keywords—Body Sensor Networks; MAC; Health Monitoring Systems; Internet of Things; Health Cloud; wireless biomedical sensor; wearable sensors; Energy optimization; motion-powered piezoelectric effect; Power consumption; time synchronization; Bandwidth utilization; memory; distributed storage; key management

I. INTRODUCTION

Physicians across the world are over loaded with vast amount of information from multi-sources, there are conflicting, incomplete diagnoses stories written and published. Most of this published material cannot really be reproduced and advancements in medical technology and patient health care remain a hit and trial and according to (WHO), World Health Organization, as many as one in ten patients in developed nation is harmed rather than cured in hospitals. All this can change for good, if continuous monitoring is done of patients using wireless network that help in not just clinical collaboration [10] but in responding in real time, on real time [6] patient's vital statistics. Progressive health care organizations are now adopting a policy of sharing, communicating in real time not just with other organizations but with the patient itself. The patient may be stationed at his/her home, office or is present in hospital itself. Now with the help of body sensor networks [1][2][4][5][19][23], heat stress [1],heart pulse [7][31], heartbeat [31], blood pressure [1][7][10] [15][20][22][31], blood sugar readings, sleep data[6][28][29] can be transmitted, and analyzed at multi-locations for fast response. This wireless body sensor network, however are connected to 'Health Cloud [32]' for remote processing by computer algorithms and healthcare service providers. The wireless biomedical sensor forms a wireless body sensor network around human body enabling patient's vital signs to be collected and transmitted in a context-sensitive manner and same is applicable which observe health of animals[15], birds[15] etc. The data stream of these vital signs are stored in distributed storage systems like Hadoop[33] and are processed in Health cloud using algorithms that detects abnormality in them. However, like in all industries, there are challenges and issues which need to be taken care for it to be successful .The systems must be able to match health data compliance, privacy issues, secure connectivity with event synchronization and problems related to its rate of changes and size /volume of the data. There are many options that offer solutions to these problems and address the issues. These solutions include use of cloud based secure connectivity with backend storage solutions like Hadoop, this paper discusses further in this context in coming sections .

The paper discusses the issues and problems associated related to human body sensor network. The related work section discusses various aspects of Body sensor networks from basics to its construction and challenges faced by the industry. A tabular summary of problems is also given after that and last but not least discussion and future directions.

II. RELATED WORK

In this section, a systematic study is discussed in area of body sensor networks(BAN), the objective of this study is to arrive at, stage of learning where progress iterations done in this are covered for identification of the problems , issues, applications, algorithms involved in building body sensor networks. A network that can be build based on wearable sensors, implantable.

This research paper envisions collection of neurological high dimension; high resolution for brain imaging that is put for use in medical for mankind. The researchers are aiming here to build a toolkit that supports brain imaging data collection and analysis the brain process. The paper also throws light on other toolkits available in this area ;(DETECT, EEGVIS, BCLLAB).It also discusses limitations of contempory efforts done in this context with respect to hardware. Limitations of high bandwidth requirement for building such long term, wearable systems of data collection from wireless networks attached to brain are discussed.

The paper points out, all though lot of research is happening in this area but many of these lack richness and depth of real world neuroimaging ,hence their research is significant in this direction especially in direction of artifact like eyes movement (left blink, right blink),head rotation , jaw movements, etc. The paper also discusses the issue of developing minimal intrusive techniques/methods in detail.

[1] "Leveraging Knowledge From Physiological Data: On-Body Heat Stress Risk Prediction With Sensor Networks" This research work have focused on wearable sensors and have trained the empirical data generated by sensors with Bayesian net algorithm and decision tree and have achieved accuracy 92.1±2.91 and 94.4±.The algorithm here, basically is predicting heat stress,CO2 level using body sensor network. Since, this algorithm is based on probability concepts the values of initial probability are provided and then based on the various conditions the value of probability is computed to arrive at the accuracy level. The graph (Figure 1) below shows skin temperature sample collection data using EOD suit.

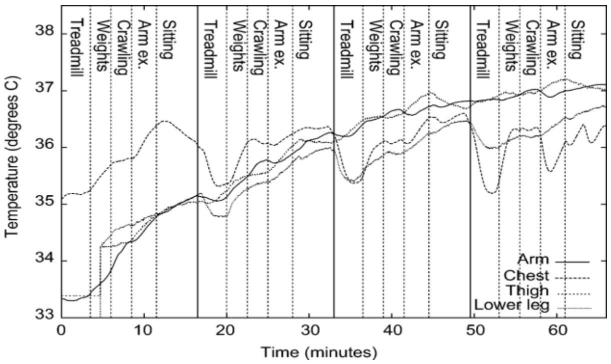
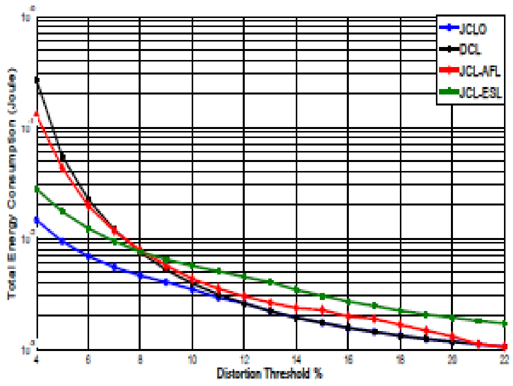
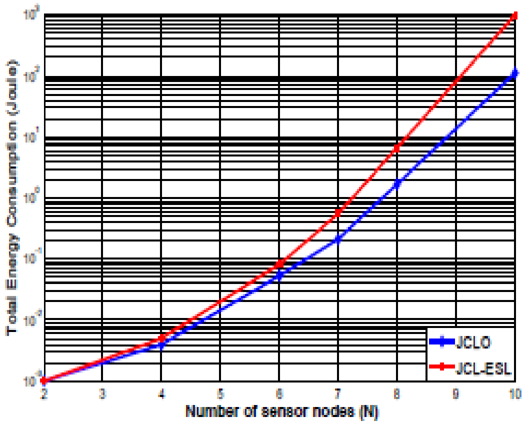


Fig. 1. Skin temperature data for a sample subject gathered during a mission like protocol while wearing an EOD suit. [1]

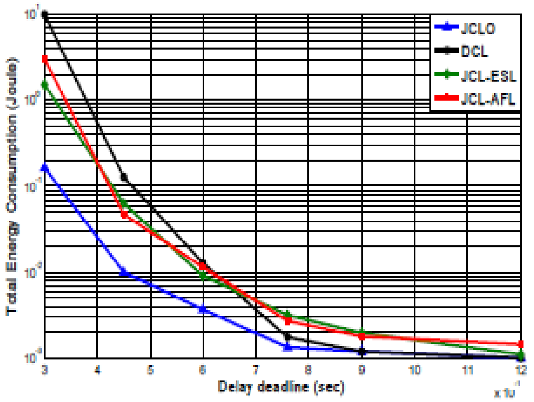
[4] "Energy-aware cross-layer optimization for EEG-based wireless monitoring applications" In this research paper energy optimization algorithm has been explained. The researchers have tried formulate an objective function that works to find an optimal tradeoff between the various design parameters along with constrains. The real idea is to minimize the total energy consumption subject to controls including data delay deadlines and distortion. The work is here about cross layer (Application-MAC-Physical). The result shows a delay deadlines increases energy consumption decreases. In graphs (Figure 2) below shows energy consumption patterns demonstrated in this research paper.



(a)



(b)



(c)

Fig. 2. A comparison between the total energy consumption using the proposed algorithms(a) For different distortion thresholds, (b) With increasing the number of sensor nodes and (c) For different delay deadlines[4]

[5] "A Motion-Powered Piezoelectric Pulse Generator for Wireless Sensing via FM Transmission" This paper demonstrates a process of building a motion-powered piezoelectric effect pulse generate for wireless sensing that using FM transmission can become a part of body sensor monitoring system. This device basically helps in energy harvesting attached with body of a subject. The paper shows engineering designs as well as mathematical relationship between the parameters that impact the performance of device like motion etc

[6] "Wireless Wearable Multisensory Suite and Real-Time Prediction of Obstructive Sleep Apnea Episodes" In this research work obstructive sleep apnea (OSA) dataset is used for classification using direct process based mixture Gaussian process (DPMG). The algorithm based can work for all body network sensors that can capture sleep data, from which a set of observation sleep apnea can be detected. The accuracy of this algorithm is 87%. This paper demonstrates that how a supervised learning algorithm when accurately grouped can result in correctly predicted Obstructive Sleep Apnea episodes of a subject under observations. The bar graph (Figure 3) below shows the values of statistical variation of feature distribution of non and apnea cases.

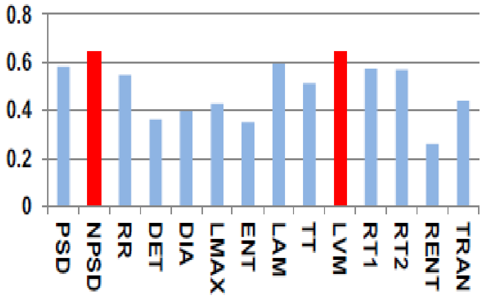


Fig. 3. KS Statistic variations of extracted feautures,KS ststistic indicates the maximal feature distribution differences between sleep apnea and non-apnea.[6]

[7] "A Neo-Reflective Wrist Pulse Oximeter" This paper illustrates wrist pulse oximeter building process with its performance evaluation. The design has 2 LED, or a receiver to capture wrist pulse signals from inner part of the wrist. It is improved version, true oximeter as compared to usual finger pulse oximeter. The calibration process shows second order relationship between 'R' value and oxygen saturation can provide better fit with the data , this is also illustrated in graph(Figure 4) below . The methods demonstrated here are based on linear and quadratic equation model of data fitting , and as per the graph values the curve made by quadratic equation model is doing the data fitting better as it is following non linear curve which is closer to the real life situation . This paper also helps us understand the internal working of body sensors in brevity.

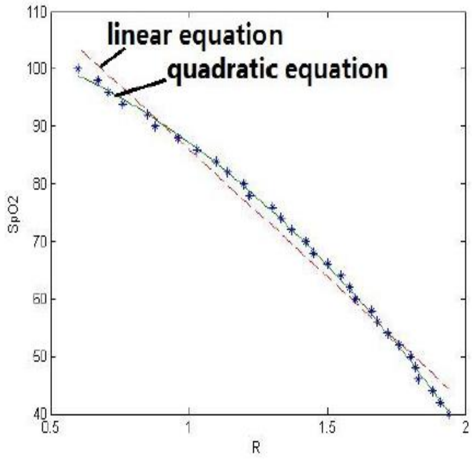


Fig. 4. The R curve modelled by two relationships[7]

[8] "An Attachable Clothing Sensor System for Measuring Knee Joint Angles" This research work shows that flexible natural based sensors can be incorporated in clothing, for measuring knee join angles. This research work was done on 10 subjects and the device showed only 1 of errors in measurements of angle. The (figure 5) shows the stress strain relationship captured using Knee Angle Sensor demonstrated in this research paper. From the graph it can be learned that initially the strain is changing faster into higher values when the strain is less < 0.5. After this it moves higher in value but at slow rate and find it move up to marginally come down by few points.

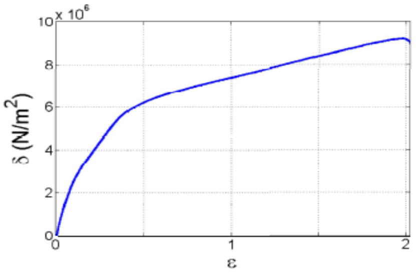


Fig. 5. Stress-strain curve obtained from the ACS sensor[8]

[9] "A Noncontact Capacitive Sensing System for Recognizing Locomotion Modes of Transtibial Amputees" This work shows usability of locomotion mode recognition system based on electromyography. The transtibial amputees with various levels of amputation have used for conducting evaluation experiments. Finally, the data is subjected to phase dependent quadratic and accuracy between 94-96% was obtained.

[10] "Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors," This research work focused on demonstrating a combined system that monitors patient subjects using stimuli data collected from wearable sensors with clinical observations. The data is then processed with machine learning algorithms like probalistic training schemes like GMM etc.The results claimed in the paper show that these algorithms have high accuracy and true positive rates. However, the histograph below (figure 6) shows the frequency of two hundred patients with respect to their stay for the clinical study. Most of the patients stayed less than 10 days for the study and few more than 30 days.

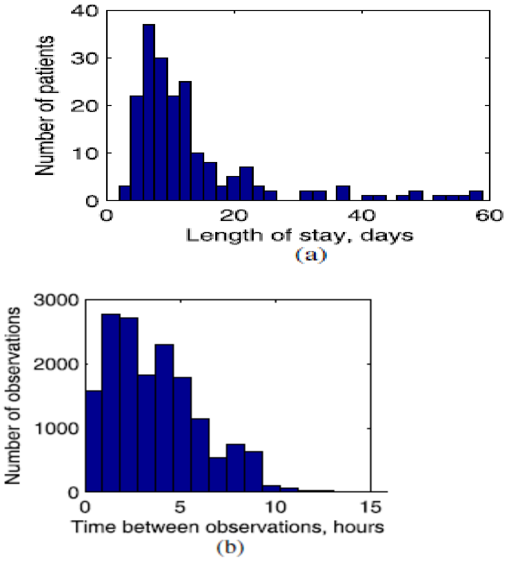


Fig. 6. (a) Histogram of the length-of-stay of 200 studied patients in the Cancer Centre. (b) Histogram of time between manual observations, over all patients[10]

[12] "A Wearable Wideband Circularly Polarized Textile Antenna for Effective Power Transmission on a Wirelessly-

Powered Sensor Platform, " This research paper shows simulation and experimental results of wideband wearable circulatory polarized textile antenna for low power transmission and it is battery less temperature sensor system that can be put in arm body. The results claimed in this paper show that the antenna performs really well with gain of 4.9Dblc.

[13] "Energy-Efficient Real-Time Human Mobility State Classification Using Smartphone," This research work tries to take advantage of the smart phone accelerometer sensors for human mobility analysis. The researcher are showing a framework based on a probabilistic algorithm that neutralizes the effect of different smart phones on body placements and orientation to allow human movements to be more accurately and energy efficiently identified. The method illustrated in this paper shows an accuracy of 92% when evaluated on a dataset, made of 15 subjects under study having a different urban mobility states like sit, stand, walk etc.

[15] "An Ingestible Sensor for Measuring Medication Adherence," This paper discusses a category of sensors that can be integrated with the oral dose of subjects and the sensors type is known as ingestible sensors. The sensors can be used for drug delivery, drug monitoring and to study drug efficiency. These sensors are micro/nano fabricated integrated circuits basically and the most important function they perform is to check medical adherence to other physiological metrics.

[16] "An adaptive home-use robotic rehabilitation system for the upper body," This paper introduces robotic prototype that mimics changes in subject is elbow angle in real-time, evaluation results show neural network based algorithm really performs well in this context of movement as error in movement, calibration etc is very low. In graph shown below the joint positions of the moving object are recorded. It can be seen that there is repetition of the movement positions, and as the graph (figure 7) is recorded the movement becomes smooth show rehabilitation.

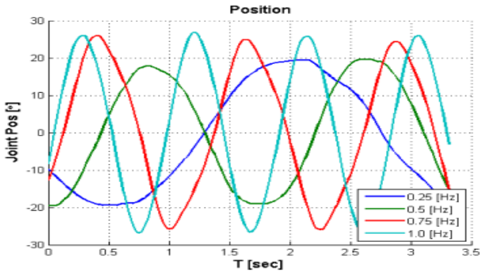


Fig. 7. Robot prototype joint angle by reference signal frequency[16]

[17] "EEG seizure detection and prediction algorithms: a survey" This paper discusses some state of art brain data seizure detection and prediction algorithms. The paper attempts to overcome the challenge of locating the seizure period in ECG recording which is done for analyzing epilepsy .The paper discusses 6 methods of seizure detection and predictions including time domain, wavelet domain, frequency domain, FCO and ICA domain. Empirical mode decomposition, summary the researcher have given a table which talks about methods, domain algorithms like time detection and prediction. Multi/single/channel database self-recorded data, frame length, features used classifiers performance methods.

[18] "An advanced physiological data logger for medical imaging applications." The paper discusses the advances in physiological data logging technology (wired and wireless) for medical imaging applications, which involve use of microSD, USB, Bluetooth wireless technologies. The equipment demonstrated here can be used in wireless capsule endoscopy and skin temperature logging applications. The paper also discusses how images are captured which need to be processed to finally be logged for useful medical purposes.

[19][20] “A power efficient MAC protocol for wireless body area networks” These papers discusses MAC level protocol for wireless body sensor networks.The objective of this protocol for increase life time by reducing idle state power consumption and increase sleep time of a BAN node. The research claim this protocol is applicable for implanted, wearable sensor devices where data rate may high as in case of mp4 streaming or medium like in case of endoscope capsule.

[21] “Experimental characterizations of a UWB channel for body area networks “This paper discuses ultra wideband (UVB) technology that is used for wireless body area network. The results claimed in this research show that human body does not significantly degrade the impedance of monopole antenna. The measured path loss and multipath analysis suggest that this technology is excellent for low power, high rate transmission applications.

[22] "Modeling on-body dtn packet routing delay in the presence of postural disconnections” This paper discusses a stochastic framework for packet routing in body sensor networks. Delay modeling method is experimented, evaluated via simulation. The results show that on body sensor count can be safely be reduced without packet delivery delay due to store and packet routing mechanism depicted in this paper. This research work has also demonstrated the performance of the body sensor network in terms of packet delivery ratio, delay etc., which is graphically shown in the graphs below. The graphs (figure 8 & figure 9) show comparative few of six algorithms.

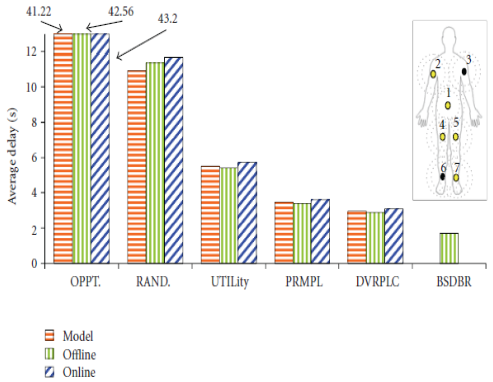


Fig. 8. On-body delivery delay for different DTN routing protocols[22]

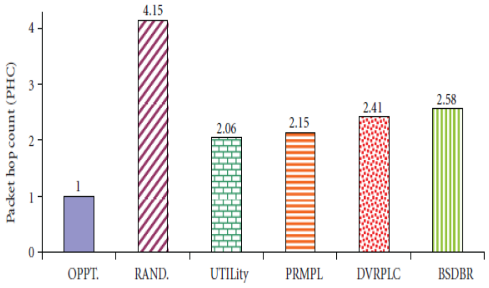


Fig. 9. Average packet hop count[22]

[23] "Design and analysis of an energy-saving distributed mac mechanism for wireless body sensor networks." This paper discusses the performance of distributed MAC process (DMAC) for wireless body sensor networks; this is done by incorporation of energy aware radio activation policies into the high performance of distributed queuing medium. The paper shows that energy assumption per bit of information vs relative traffic load increase slowly initially and later in after 0.7 value of relative traffic load increases exponentially as in (figure 10)

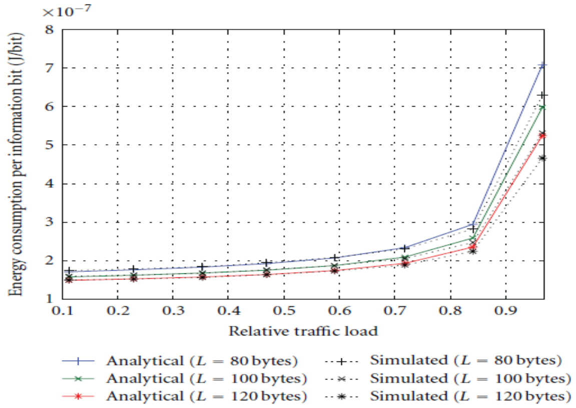


Fig. 10. DQ-MAC energy consumption per information bit:Analytical versus Simulation[23]

[26] “On PHY and MAC performance in body sensor networks.” In this research work an empirical study has been conducted to evaluate the performance of body sensor implement communicating using radio. This paper tries to address the first concern which a person may have, which is observation is one thing, with body sensor device implanted in human body may lead to some side effects on the well being of a person, especially effect in body tissues. The researchers found that best performance of the sensor comes at 3cm depth inside the liquid and not close to skin surface. Later section of this paper also shows simulation study of many low powered MAC protocols and cooperative view on them. In the graphs shown below, we can infer the relationship between the ERP and depth of liquid and RSSI and depth. But these graphs help us understand effectiveness of the sensor with respect to the depth of skin where the sensor is applied which can be validated from the behavior of ECC average coverage as shown in (figure 11,12,13)

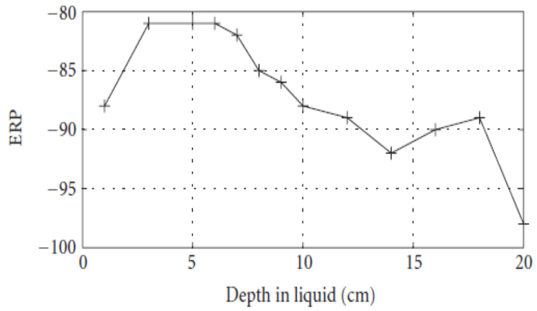


Fig. 11. ERP versus Depth in liquid[26]

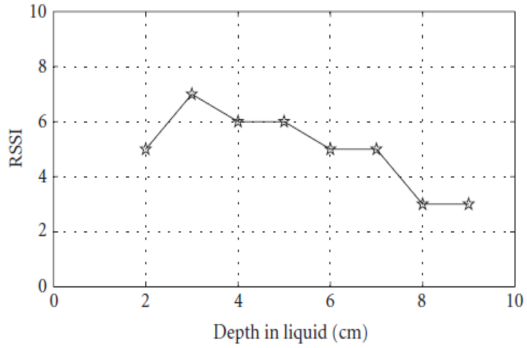


Fig. 12. RSSI versus Depth in liquid[26]

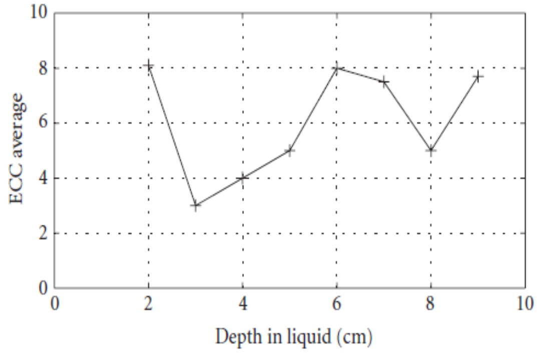


Fig. 13. ECC invocation versus Depth in liquid[26]

[30] "Biometric methods for secure communications in body sensor networks: resource-efficient key management and signal-level data scrambling." The paper highlights the challenges in developing resource efficient key management and signal level data scrambling for making it more secure. The researchers have used ECG signal for their study. The resulting simulation shows that efficacy increases with implementation of these key fusion methods.

III. Main Issues and Challenges

Based on the above systematic review of papers, materials associated with multiple aspects of emerging technologies in context of body sensors.
We can identify following gaps, limitations of existing work done in this area as shown in Table No I. and illustrated in Figure 14.

TABLE I. MAIN ISSUES AND CHALLENGES

S.NO	**Main Issues And Challenges**	
	Parameter	***Challenges***
1.	Power Consumption[27]	The requirement is ultra low-powered devices.
2.	Sleep Cycle , Wakeup periods	Synchronization of sleep and cycles with stimuli and with full body sensor network.
3.	Idle listening time	This again is issue of synchronization of all devices as per stimuli.
4.	Overhead(Control packet)	If the application of BAN is specific to particular type of stimuli. There is no need for elaborate protocol .This way we can reduce protocol stack size there by overheads.
5.	Packet collision, Retransmission[21][26]	If sudden outburst of stimuli. There is large volume of network traffic that must flow smoothly so that packets reach without delay , without collisions , retransmissions .It become challenge as routing soft components are frugal in case of WBAN .
6.	Bandwidth utilization[21][26]	Most of devices can handle small bytes, this limits the to and fro flow of data. Hence, effective use Bandwidth is critical for such devices .
7.	Seizures of stimuli[17][2]	Loading, recording the seizures of stimuli are very challenging in case of brain and when body is in movement.
8.	Storage of data collection, Time lag[18][3][6]	The need is distributed storage.
9.	Memory	Limited to store and process.
10.	Signal Integrity	Signal must maintain its coherence, shape, energy and spectrum properties and no agent must interfere /manipulate it .
11.	Signal scrambling	This can really secure signal, but at the same time this may lead to overload in terms of delay and reduced synchronization.
12.	Signal shape	No interference, no noise or distortion should be there when signal reaches control room.
13.	Key management , Key generation , Key Exchange , Key recordability ,Key [27][30][22]	Operational resources in body sensor networks are highly restricted, incorporation of key management scheme lead to overhead, synchronization issues at the lost of security which is also essential.
14.	Error in measurement of sensor[8]	This depends upon the accuracy of digital instrument tool with which measurements are taken both in terms of hardware and software components.
15.	Area, Volume and Weight [12][19][20][23]	Unless nano, micro size is realized, there is limited scope of development in this area, as a human body must not feel burden of sensors in terms of weight, area ratio and volume.
16.	Harmful effects of body sensors on human tissues and overall well being	This technology‘s side effects still need to be observed and recorded for understanding its ill effects on human body.
17.	Inert and Green technology[24]	Body sensors must not react with human tissues and affect the health due to radiation transmission of signal waves.
18.	Integration of sensor data with other clinical data.[10]	This is critical for proper working and spirit of the concept of continuous monitoring so that physician can take diagnostic decisions.
19.	Integration to main stream medical technology[10]	Body sensors network data must be interoperable for it to be used across departments and must be able to follow standards like HL7 etc.
20.	Time Synchronization[14][22]	Final transmission to control room may lead to problems, challenges in terms of delay, time lag and signal may require pre and post process.
21.	Detection	In built anomaly detection, adversity identification systems are must, as these may prone to DDOS, Sybil etc, attacks.
22.	Prediction	Based on historical data, WBAN must have algorithm to predict health issues like heart attack [31] for its real success and application.

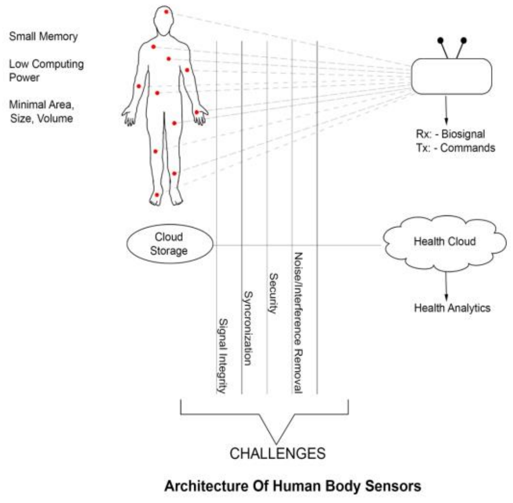


Fig. 14. Architecture of Human Body Sensors

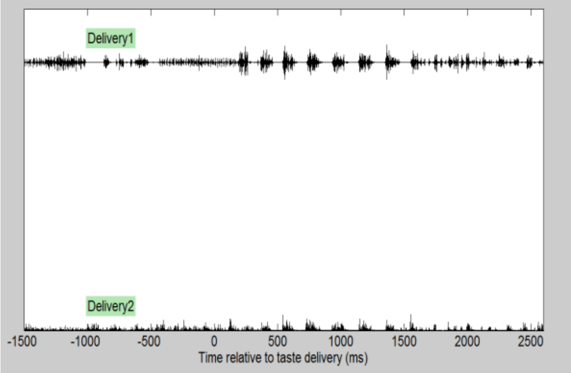


Fig. 15. Example of Human Body Sensors picking up Drug & Taste stimuli[34]

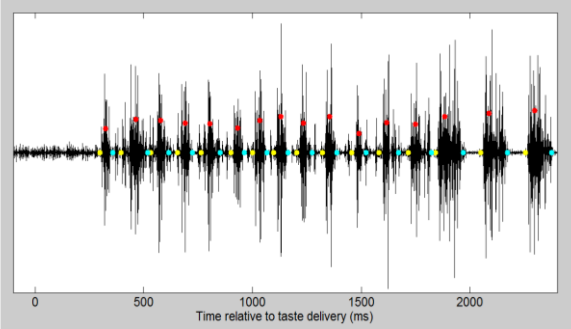


Fig. 16. Example of Human Body Sensors picking up Drug & Taste stimuli[34]

The Figure 15 & Figure 16 shows the EMG spike response when a drug shows its effect on the human subject and at the same time it shows the taste based body reaction /stimuli when the drug is tasted from brain captured by human body sensors.

IV. Conclusion and Discussion

The applications are enormous of body sensor networks. They can be used right from the birth of a person till his/her old age.

Body sensor networks are now been developed for disease management and previous oriented healthcare. These systems may be synchronous or asynchronous in nature with multi-body platform designs and implementation may help detect motor patterns, heat stress for example. Now, these networks are also now being experimented in understanding group dynamism like understanding vital signs of a cricket team. However, it should also be noted that, some of the implication of the problems are quite grave, if link interruptions or failures occurs in “internet of medical equipment” it can lead to unwanted consequences and other medical system complications.

Then, there are striking implications in case of data ownership in case we are using some fitness trackers and medical devices. Who should own the data related to the heartbeats? Person, community or some fitness device company.Even if the answer would seem obvious, but in today’s environment it’s not so clear. Existing technology in body network are not even addressing existing problems technically , and further development is likely to exacerbate that state of affairs due to inherent nature of Body sensor network technology.

In this research work, we have done systematic study for identification of applications, implications of the problems development in area of body sensor networks. The paper attempts to cover all that would impact the building, deployment and working of the body sensors in technical sense. The paper also covers MAC, physical layer issues as well as other hardware components, building locks used in build antennas for body sensors.

However, we have not covered discussion on the ethical issues related to Human Body sensors although, future similar work may cover up moral and ethical issues also.

V. Future Direction in Body Area Network

Future of WBAN lies in volume of implementations, installation and adoption of this technology by people at large. There is huge development, experimentation going on this context. In no time, these networks will become important tool in today’s health. However, the challenges are also there, especially related to secure transmission and integrity of signals. For future directions, we suggest, artificial algorithms must be used for securing signal shapes and communication as well maintain integrity.

Acknowledgment

I would like to take this opportunity to thank my guide Manik Gupta, and my parents for helping me in the research study.

References

[1] Gaura, E., Kemp, J., Brusey, J., "Leveraging Knowledge From Physiological Data: On-Body Heat Stress Risk Prediction With Sensor Networks," Biomedical Circuits and Systems, vol. 7, pp. 861-870, Dec. 2013.

[2] McDowell, K., Chin-Teng Lin, Oie, K.S., Tzyy-Ping Jung; Gordon, S.Whitaker, K.W., Shih-Yu Li, Shao-Wei Lu, Hairston, W.D., "Real-World Neuroimaging Technologies," IEEE, vol. 1, pp. 131-149, 2013.

[3] Costlow, T., "Camera phone bans expected," Distributed Systems Online, vol. 5, pp. 5/1-5/3, Feb. 2004.

[4] Awad, A., Hussein, R., Mohamed, A., El-Sherif, A.A., "Energy-aware cross-layer optimization for EEG-based wireless monitoring applications," Local Computer Networks (LCN), 2013 IEEE 38th Conference , pp. 356-363, Oct. 2013

[5] Hao Jiang, Kiziroglou, M.E., Yates, D.C., Yeatman, E.M., "A Motion-Powered Piezoelectric Pulse Generator for Wireless Sensing via FM Transmission," Internet of Things Journal, IEEE , vol. 2, pp. 5-13, Feb. 2015

[6] Le, T.Q. ,Changqing Cheng, Sangasoongsong, A. , Wongdhamma, W.,Bukkapatnam, S.T.S., "Wireless Wearable Multisensory Suite and Real-Time Prediction of Obstructive Sleep Apnea Episodes," Translational Engineering in Health and Medicine, IEEE Journal, vol. 1, pp. 2700109-2700109, 2013

[7] Pang, G., Chao Ma, "A Neo-Reflective Wrist Pulse Oximeter," Access, IEEE , vol. 2, pp. 1562-1567, 2014

[8] Bergmann, J.H.M., Anastasova-Ivanova, S., Spulber, I., Gulati, V., Georgiou, P., McGregor, A., "An Attachable Clothing Sensor System for Measuring Knee Joint Angles," Sensors Journal, IEEE , vol. 13, pp.4090-4097, Oct. 2013

[9] Enhao Zheng, Long Wang; Kunlin Wei, Qining Wang, "A Noncontact Capacitive Sensing System for Recognizing Locomotion Modes of Transtibial Amputees," Biomedical Engineering, IEEE, vol. 61, pp. 2911-2920, Dec. 2014

[10] Clifton, L., Clifton, D.A., Pimentel, M.A.F., Watkinson, P.J.,Tarassenko, L., "Predictive Monitoring of Mobile Patients by Combining Clinical Observations With Data From Wearable Sensors," Biomedical and Health Informatics, IEEE Journal, vol. 18, pp. 722-730, May 2014

[11] Ngai, G., Stephen Cf Chan; Cheung, J.C.Y., Lau, W.W.Y., "Deploying a Wearable Computing Platform for Computing Education," Learning Technologies, IEEE , vol. 3, pp. 45-55, Jan.-March 2010

[12] Lui, K.W., Murphy, O.H., Toumazou, C., "A Wearable Wideband Circularly Polarized Textile Antenna for Effective Power Transmission on a Wirelessly-Powered Sensor Platform," Antennas and Propagation, IEEE , vol. 61, pp. 3873-3876, July 2013

[13] Oshin, T.O. ,Poslad, S., Zhang, Z., "Energy-Efficient Real-Time Human Mobility State Classification Using Smartphones," Computers, IEEE , vol. PP, pp. 1-1

[14] Goth, G., "Delay-Tolerant Network Technologies Coming Together," Distributed Systems Online, IEEE , vol. 7, pp. 2,2, Aug. 2006.

[15] Hafezi, H., Robertson, T.L., Moon, G.D., Kit-Yee Au-Yeung; Zdeblick, M.J., Savage, G.M., "An Ingestible Sensor for Measuring Medication Adherence," Biomedical Engineering, IEEE , vol. 62, pp. 99-109, Jan. 2015

[16] Dowling, A.V., Barzilay, O., Lombrozo, Y., Wolf, A., "An adaptive home-use robotic rehabilitation system for the upper body," Translational Engineering in Health and Medicine, IEEE Journal, vol. 2, pp. 1-10, 2014.

[17] Alotaiby, Turkey N., Saleh A. Alshebeili, Tariq Alshawi, Ishtiaq Ahmad, and Fathi E. Abd El-Samie. "EEG seizure detection and prediction algorithms: a survey." EURASIP Journal on Advances in Signal Processing ,2014.

[18] Khan, Tareq Hasan, and Khan A. Wahid. "An advanced physiological data logger for medical imaging applications." EURASIP Journal on Embedded Systems, pp. 1-1 , 2014.

[19] Koulali, Mohammed-Amine, Abdellatif Kobbane, Mohammed El Koutbi, Hamidou Tembine, and Jalel Ben-Othman. "Dynamic power control for energy harvesting wireless multimedia sensor networks." EURASIP Journal on Wireless Communications and Networking, pp. 1-8 , 2012,.

[20] Al Ameen, Moshaddique, Niamat Ullah, M. Sanaullah Chowdhury, SM Riazul Islam, and Kyungsup Kwak. "A power efficient MAC protocol for wireless body area networks." EURASIP Journal on Wireless Communications and Networking, pp. 1-17, 2012.

[21] Xia, Lingli, Stephen Redfield, and Patrick Chiang. "Experimental characterization of a UWB channel for body area networks." EURASIP Journal on Wireless Communications and Networking , 2011.

[22] Quwaider, Muhannad, Mahmoud Taghizadeh, and Subir Biswas. "Modeling on-body dtn packet routing delay in the presence of postural disconnections." EURASIP journal on wireless communications and networking , 2011.

[23] Begonya, Otal, Alonso Luis, and Verikoukis Christos. "Design and analysis of an energy-saving distributed mac mechanism for wireless body sensor networks." EURASIP Journal on Wireless Communications and Networking , 2010.

[24] Naveen, Chilamkurti, Zeadally Sherali, Jamalipour Abbas, and Das Sajal K. "Enabling Wireless Technologies for Green Pervasive Computing." EURASIP Journal on Wireless Communications and Networking 2009 , (2010).

[25] Kolar, Anthony, Olivier Romain, Jade Ayoub, David Faura, Sylvain Viateur, Bertrand Granado, and Tarik Graba. "A system for an accurate 3D reconstruction in video endoscopy capsule." EURASIP journal on embedded systems , 2009 .

[26] Sana, Ullah, Higgins Henry, Islam SM Riazul, Khan Pervez, and Kwak Kyung Sup. "On PHY and MAC performance in body sensor networks." EURASIP Journal on Wireless Communications and Networking , 2009.

[27] Kaur, Jasdeep, and Sandeep Singh Gill. "QoS based energy efficient key management in body sensor networks." In Medical Imaging, m-Health and Emerging Communication Systems (MedCom),IEEE , pp. 14-19 , 2014.

[28] Chen, Min, Sergio Gonzalez, Athanasios Vasilakos, Huasong Cao, and Victor C. Leung. "Body area networks: A survey." Mobile networks and applications , pp. 171-193 , 2011.

[29] Latré, Benoît, Bart Braem, Ingrid Moerman, Chris Blondia, and Piet Demeester. "A survey on wireless body area networks." Wireless Networks , pp. 1-18 , 2011.

[30] Bui , Francis Minhthang , Dimitrios Hatzinakos. "Biometric methods for secure communications in body sensor networks: Resource-efficient key management and signal-level data scrambling." EURASIP Journal on Advances in Signal Processing , 2008

[31] Huaming Li, Student Member, IEEE, and Jindong Tan, Member, IEEE" Heartbeat-Driven Medium-Access Controlfor Body Sensor Networks" IEEE transactions on information technology in biomedicine, vol. 14, Jan. 2010

[32] Misra, S., Das, S.; Khatua, M., Obaidat, M.S. "QoS-Guaranteed Bandwidth Shifting and Redistribution in Mobile Cloud Environment", Cloud Computing, IEEE , vol. 2 , pp. 181 - 193 , April-June 2014

[33] Sobhy, D. , Coll. of Comput. Eng., Arab Acad. of Sci. & Technol. & Maritime Transp., Alexandria, Egypt , El-Sonbaty, Y. , Abou Elnasr, M. "MedCloud: Healthcare cloud computing system" IEEE , pp. 161 - 166, Dec. 2012.

[34] Li, Jennifer X., Takashi Yoshida, Kevin J. Monk, and Donald B. Katz. "Lateral hypothalamus contains two types of palatability-related taste responses with distinct dynamics." The Journal of Neuroscience 33, pp. 9462-9473, 2013.

Vague Set Theory for Profit Pattern and Decision Making in Uncertain Data

Vivek Badhe
Department of Computer Applications
MANIT, Bhopal, India

Dr. R.S Thakur
Department of Computer Applications
MANIT, Bhopal, India

Dr. G.S Thakur
Department of Computer Applications
MANIT, Bhopal, India

Abstract—**Problem of decision making, especially in financial issues is a crucial task in every business. Profit Pattern mining hit the target but this job is found very difficult when it is depends on the imprecise and vague environment, which is frequent in recent years. The concept of vague association rule is novel way to address this difficulty. Merely few researches have been carried out in association rule mining using vague set theory. The general approaches to association rule mining focus on inducting rule by using correlation among data and finding frequent occurring patterns. In the past years *data mining* technology follows traditional approach that offers only statistical analysis and discovers rules. The main technique uses support and confidence measures for generating rules. But since the data have become more complex today, it's a requisite to find solution that deals with such problems. There are certain constructive approaches that have already reform the ARM. In this paper, we apply concept of vague set theory and related properties for profit patterns and its application to the commercial management to deal with Business decision making problem.**

Keywords—Association Rule Mining; Vague Association Rule Mining; Profit Pattern Mining

I. Introduction

Pattern discovery from huge volume of data is one of the most desired attributes of Data Mining [5]. However, in reality, a substantial portion of the available information is stored in text databases, which consists of large collections of documents from various sources, such as news articles, books, digital libraries and Web pages. Since web search engines have become pervasive and search has become integrated, retrieving of information from these search engines consist of three essentials: query, documents, and search results.

The emerging growth of data mining raises the large range of complex applications [14]. It leads the broad study of data mining frequent patterns. Mining frequent sets over data streams present attractive new challenges over traditional mining in static databases. Data mining is generally used for retrieving the desire information to make it into knowledge from the large size databases.

The study shows [16] that interestingness measures are distinct for different applications and substantiate that domain knowledge is necessary to the selection of an appropriate interestingness measure for a particular assignment and business objective and the goal of any business is to generate profit. So the profit can be taken as one of the measures with proper mining technique so it can help in decision making process of business.

The rest of the paper is organized as follows, in section 2 we discuss the fundamental basis of association rules, and vague set theory that deal with uncertainty and vagueness, Section 3 introduces the related work that has been done so far with this theory in accordance with association rules called Vague Association Rule (VAR)[10]. In Section 4 describe our methodology and discuss the How VAR is helpful to business problem and its decision making process. In section 5 discussions of result and comparison with classical and vague rules. Finally Section 6 concludes the paper.

II. Preliminaries

A. Association Rule Mining

Association rules discovery is one of the most important method which was given by R. Agrawal in 1993 [1]. It gives the information like "if-then" statements. These rules are invoked from the dataset. It generates from calculation of the support and confidence of each rule that can show the frequency of occurrence of a given rule. Association Analysis is the process of discovering hidden pattern or condition that occurs frequently together in a given dataset. Association Rule mining techniques looks for interesting associations and correlations among data set. An association rule is a rule, which entails probabilistic relationship, with the form $X \Rightarrow Y$ between sets of database attributes, where X and Y are sets of items, and $X \cap Y = \phi$. Given the set of transactions T, we are interested in generating all rules that satisfy certain constraints. These constrains are *support* and *confidence*. The *support* of the rule is the fraction of the transactions in T that satisfy the union of items in X and Y. The probability, measured as the fraction of the transactions containing X also containing Y, is called the *confidence* of the rule.

Support should not be confused with confidence. While confidence is a measure of the rule's strength, support corresponds to statistical significance.

With the help of these constraints, rules are computed from the data and, association rules are calculated with help of probability. Mining frequent itemsets [17] is a fundamental and essential problem in many data mining applications such as the discovery of association rules, strong rules, correlations, multi-dimensional patterns, and many other important discovery tasks. The first and foremost algorithm that was given to generate association rules was *apriori* [2]. Its proposal used the same two constraints: support and confidence, and forming rules in accordance with these constraints.

B. Profit Pattern Mining

Ke Wang, Senqiang Zhou, and Jiawei Han in 2002 presented a profit mining [18] approach to reduce the gap between the statistic-based pattern extraction and the value-based decision making. They took a set of past transactions and pre-selected target items, and intended to build a model for recommending target items and promotion strategies to new customers, with the goal of maximizing the net profit. They identified several issues in profit mining and proposed solutions. They evaluate the effectiveness of this approach using data sets of a wide range of characteristics. The key to profit mining is to recommend “right” items and “right” prices. If the price is too high, the customer will go away without generating any profit; if the price is too low or if the item is not profitable, the profit will not be maximized. The approach is to exploit data mining to extract the patterns for right items and right prices. The key issues in this context are Profit based patterns, shopping on unavailability, explosive search space, optimality of recommendations, and interpretability of recommendation.

C. Vague Set Theory

The classical (crisp) set theory define sets as the “collection of objects (either similar or dissimilar) called elements of a set as a whole”. These are also referred as crisp in nature because they only tells whether an element is a member or not, i.e., either 0 or 1. It may be also given as an element belongs to or does not belong to particular set. The crisp set theory often times is unable to provide a better understanding of any object/element to be of a certain group. Thus, leading to the fact that value might lie in between 0 and 1.

A *vague* set is a set of element distributed in a universe that has a grade of membership values in the continuous subinterval of [0, 1]. Hence, such a set can be marked by *true membership* and *false membership* functions. The continuous subinterval states both about the evidence that is in favor of the object and also that is opposing it.

In early 90’s Gau’s and Buehrer [4] introduced the notion of vague sets. Let V be the vague set. If U is the universe of discourse having X objects with *x* elements than V in U can be defined using the true membership (V_t) and false membership (V_f) functions. Considering that both V_t and V_f consorted as real numbers in the subinterval of [0, 1]. Also V_t is the lower bound on grade of membership of *x* derived in favor of *x*, and V_f is the lower bound on grade of membership derived against *x*, with each element in X where $V_t + V_f \leq 1$ and V_t:X→[0,1], V_f:X→[0, 1]. Hence, the grade of membership of *x* is bounded to a subinterval $[V_t(x), 1\text{-}V_f(x)]$ of [0, 1]. A vague set V in a universe of discourse U is characterized by a true membership function, t_v and a false membership function, f_v, as follows:

$$t_v\colon U \rightarrow [0, 1],$$

$$f_v\colon U \rightarrow [0, 1], \text{ and}$$

$$t_v(x) + f_v(x) \leq 1,$$

where $t_v(x)$ is a *lower bound* on the grade of membership of u derived from the evidence for *x*, and $f_v(x)$ is a lower bound on negation of *x* derived from the evidence against *x*. An Lu and Wilfred Ng [9] gave a detailed discussion of using the proper theory for imprecise or vague data.

D. Vague Association Rules

The notion of *Vague Association Rules* (*VARs*)[10] is based on four types of support and confidence, which applied on AH-Pair Transactions Given the transactions of the customers, we then aggregate the transactions to obtain the *intent* of each item. Based on the intent of an item, we next define the *attractiveness* and *hesitation* of it [6,7,8] .

(Intent, Attractiveness and Hesitation, AH-**Pair Transactions)** The intent of an item x, denoted as intent(x), is a vague value [α(x), 1 − β(x)]. The attractiveness of x, denoted as M_A (x), is defined as the median membership of x, i.e., $M_A(x) = (\alpha(x) + (1 - \beta(x)))/2$. The hesitation of x, denoted as M_H (x), is de-fined as the imprecision membership of x, i.e., $M_H(x) = ((1 - \beta(x)) - \alpha(x))$. The pair $hM_A(x), M_H(x)i$ is called the AH-pair of x. An AH-pair transaction T is a tu-ple $<v_1, v_2, \ldots, v_m>$ on an itemset $I_T = \{x_1, x_2, \ldots, x_m\}$, where $I_T \subseteq I$ and $v_j = hM_A(x_j), M_H(x_j)i$ is an AH-pair of the item x_j, for $1 \leq j \leq m$. An AH-pair database is a sequence of AH-pair transactions.

(Vague Association Rule) *A* Vague Association Rule (VAR), r = (X => Y), *is an association rule obtained from an AH-pair database.*

Based on the attractiveness and hesitation of an item, we define four different types of support and confidence of a VAR depending on what kind of knowledge we want to acquire. For clarity, we use A to denote *Attractiveness* and H to denote *Hesitation*.

Support and Confidence

Given an AH- pair database, D , we can define different types of support and Confidence for an itemset Z or a VAR $_X \Rightarrow {}_Y$,where $X \cup Y = Z$ [6].

$$\text{Support} \quad T_S \quad = \sum \text{Mp} / |D|$$

Where

T_S = {A-sup, H-Supp,AH-Supp,HA-Supp}
p = {X,Y,XUY}
$M = \{M_A,M_H, M_A.M_H\ M_HM_{A\}}$

$$\text{Confidence C} = T_S(Z)/\ T_S(X)$$

E. Uncertainty in Data Mining

The traditional data mining approach uses statistical and logical significance to find the knowledge from the databases. Since the databases have become diverse and heterogeneous which contain data close to real world, they are susceptible to uncertainty. By uncertainty we mean that it is not possible to depict the true nature of the data and what will be the outcome of it when processed. Uncertainty occurs when it is impossible to assert any value to an object when modeling is done [9,10]. Uncertainty can be of distinct forms and to identify specific one is a taxing work. Some types of uncertainties are:

- *Imprecision*: the available information is not specific to the desired modeling.
- *Inconsistency*: there are two or more statements in modeling which cannot be true at same instant.
- *Ambiguity*: the objects in the model have stringency because of which many possible renditions can be made.
- *Vagueness*: the objects in a model include an intrinsic vague value which is not expressed clearly. Vagueness is formalized from the concept of fuzziness.

To deal with uncertainty in mining, some soft computing techniques must be incorporated which helps to reason with the databases. Neural Networks, Fuzzy logic, Genetic Algorithm, Rough Sets, Vague Sets are some of the soft computing techniques that does deal uncertainty to some extent].

III. Related Work

In ARM, support and confidence are the basic measures that have been used since its inception, which define the statistical significance of any rule [16].

Sandhu, P.S. et. al. in 2010 [15] proposed an efficient approach based on weight factor and utility for effectual mining of significant association rules. Initially, the proposed approach makes use of the traditional Apriori algorithm to generate a set of association rules from a database. The proposed approach exploits the anti-monotone property of the Apriori algorithm, which states that for a k-itemset to be frequent all (k-1) subsets of this itemset also have to be frequent. Subsequently, the set of association rules mined are subjected to weight age (W-gain) and utility (U-gain) constraints, and for every association rule mined, a combined Utility Weighted Score (UW-Score) is computed. Ultimately, they determined a subset of valuable association rules based on the UW-Score computed. The experimental results demonstrate the effectiveness of the proposed approach in generating high utility association rules that can be lucratively applied for business development

An Lu and Wilfred Ng [10] provided another 4 support and 4 confidence measures based on vague properties which assists in finding more interesting rules. Merely few researches have been carried out in association rule mining using vague set theory [3].

In 2007, An Lu and Wilfred Ng[10] apply the vague set theory to address a limitation in traditional AR mining problem, that is, the hesitation information of items is not considered. They propose the notion of VARs that incorporates the hesitation information of items into ARs. They also define different types of support and confidence for VARs in order to evaluate the quality of the VARs for different purposes. An efficient algorithm is proposed to mine the VARs.

An Lu et al. [11] Modeled hesitation information by the purchaser in online shopping using Vague Association rule and provide the notion of VAR for almost sold items in the online shopping by considering different user preference.

Anjna Pandey et al. [12,13] developed the models for hesitation information for course information using vague set theory in order to address a limitation in traditional association rule mining problem, which ignores the hesitation information of items in transactions. The efficient algorithm for mining vague association rule that discovers the hesitation information of items is proposed to solve the course information, attendance and related vagueness. They extend their work for temporal Association rule mining that can be used to evaluate the course effectiveness and helps to look for in regards to changes in performance of the course from time to time.

IV. Proposed Methodology

The proposed methodology is especially developed for commercial transactions where each item having an item code but for the different packing the code is differ for the same item. This will create the vagueness and cannot be deal by the conventional association rule mining for this purpose we use the vague association rule mining and generate those rules which generate the profit significance but ruled out due to statically violation caused by vagueness. We use an algorithm to mine Vague set based Association Rules. As discussed in previous section, the vague sets have found application in association rule mining in many ways. We propose another important methodology to find support and confidence for mining association rules. The technique we propose consist three new formulas for finding support and confidence.

Definition 1: Variation Table/Matrix

The variation table matrix is a table that is formed of vague items. The variation table contains N rows depicting the number of vague items and M columns corresponding to the variant of that vague item.

Definition 2: Vague Percentage

The vague percentage denotes the amount of vagueness contained in a database for a particular vague item. For an item in a database, there exists a true membership (V_t), a false membership (V_f), and a certain amount of vagueness, as in our case vague percentage (V_p). Thus, a database consists of $|D| = V_t + V_f + V_p \leq 1$ all the memberships and percentage values. The vague percentage can be denoted now as $V_p = |D| - (V_t + V_f)$.

Definition 3: True Support (Sup_{tr})

Let A be a vague item that has A_1, A_2… as vague values, then true support Sup_{tr} is defined as $Sup_{\mathrm{tr}} = \frac{\mathrm{V_t(A_1)}}{[\mathrm{V_t(A_1)+V_f(A_1)}]}$ i.e. the true membership of an item A_1 is in ratio with the sum of its own true membership and its false membership. We do this because the database as a whole contains vagueness, which means $|D| = V_t + V_f + V_p \leq 1$. Hence on excluding vagueness from the database $(V_t + V_f \leq 1 - V_p)$ then it gives the true support of that item which classical method did not consider. The value of Sup_{tr} can also be defined in terms of Vague Percentage as $\frac{\left\{V_t + \left[\frac{V_P * V_t}{(V_t + V_f)}\right]\right\}}{|D|}$

Definition 4: True Confidence ($Conf_{tr}$)

True confidence is the ratio of true support of the union of items A and B to the true supports of any of the item either A of B. It is denoted by $Conf_{tr} = \frac{Sup_{tr}(A\cup B)}{Sup_{tr}(A)}$.

Algorithm: Vague Itemset minEr (VIE)

A. Vague Table Generation Algorithm

Variation Table : VagueTab (D)

```
1) Scan the database D to find the total number of distinct items;
2) For i = 0, 1, 2, …where i = no. of transactions T in D, do
3) Initialize both true membership ($V_t$) and false membership ($V_f$) variables with zero;
4) For j = 0, 1, 2, … where j=no. of vague items in D, do
5) Increment the vague item count for i^th item and store it in the Vague Table;
6) End of for;
7) End of for;
8) Return VagueTab(t);
```

B. Vague Itemset minEr (VIE) Algorithm

VIE (D, $\boldsymbol{V_t}$)

```
1) Calculate 1st vague frequent itemset by scanning D;
2) For i=0, 1, 2, … $t_n$ transactions in D, do
3) Call VagueTab(D) and generate candidate $C_i$ and check whether the item in candidate is vague or not;
4) If an item is vague, find its variant from VagueTab and calculate true membership ($V_t$) of that item(s);
5) If an item is non-vague then directly add in to the candidate list $C_i$;
6) Perform the pruning of the list by applying true membership ($V_t$) and generate 2nd vague frequent itemset;
7) Find next vague frequent itemset till no further combinations are possible
8) End of for;
9) Return vague frequent itemset (L);
```

C. Vague Rules Generation

VagueR(sup, conf)

```
1) find the last vague frequent itemset (L);
2) generate subsets, s of the vague itemset such that $s \in S$;
3) if s is vague frequent itemset, then find the rule;
4) else omit the subset from the itemset list;
5) return VagueR(n, m);
```

The experiment was conducted on an FMCG database that contains the daily inventory of the products purchased by users. We classify items on the basis of products denoting each with a certain unique code. The vagueness is found in the database by using the Vague Table which consist all the variants of a particular product that is available in the database. This imparts vagueness in the dataset on which we apply out Vague Itemset minEr (VIE) algorithm. Since the FMCG database is very large consisting of huge number of transactions, we only report results on a sample of transactions selected at random. The number of transactions T taken into account is 10 which consists a number of products with their codes, i.e. both vague items and non-vague items. First experiment was conducted on the sample dataset with the traditional Apriori method and the results are noted in Table 1. The second experiment was conducted on the same sample dataset with the Vague Itemset minEr (VIE) algorithm and the result are noted in Table 2. The minimum threshold support was kept at 30% and minimum threshold confidence at 80%.

TABLE I. RESULT OF APRIORI

		(min_sup, min_conf)
AAT001	<- ABD022	(50, 80)
AAT001	<- ABB012	(50, 80)
AAE038	<- ABB012	(50, 80)
AAT001	<- AAE165 ABB012	(30, 100)
ABB012	<- AAE165 AAT001	(30, 100)
AAT001	<-ABD022 ABB012	(30, 100)

TABLE II. RESULT OF VAGUE CONSIDERED ARM

		(min_sup, min_conf)
ABD022	<- AAG272	(40, 82)
AAE038	<- AAE165	(40, 82)
ABB012	<- AAE165	(40, 82)
ABB012	<- AAT001	(60, 83)
AAE038	<- ABB012	(83, 96)
ABB012	<- AAE038	(85, 94)
ABB012	<- AAE165 AAT001	(30, 100)
AAT001	<- AAE165 ABB012	(33, 90)

It is clear from the Table 1 & 2 that the rules that are generated from traditional Apriori have only three rules (highlighted above for visual understanding) that are also found using VIE algorithm but with difference in their support and confidence measures. Some of the rules that are in Table 1 are omitted in Table 2 and vice versa. This demonstrates that the traditional Apriori performs undermining and forms *subversive rules* whereas the vague algorithm VIE performs over mining and forms *puissant rules*. Thus vague sets when incorporated with association rules provide a contrasting meaning to the classical approach and gives better results.

V. DISCUSSION

The classical approach to association rule mining uses the Apriori or similar algorithms which are based purely on statistical significance of the items present in the database. The approach was easy to consider when the databases contained only textual data, or in other words, certain data. As the database technology evolved, the basic measures of support and confidence were proving to be scarce in finding relevant knowledge. It is evident that without support and confidence it is difficult to find rules but to improve the knowledge discovery some more measures or dimensions need to be incorporated. One way of doing it is by first understanding our database. There are many mathematical tools that have been used over time to fulfill ones requirement with databases. The approach we propose takes uncertainty in consideration. Now, there are many types of uncertainty that could be handled each with a different and specific tool. The uncertainty we consider is of vagueness type.

Vagueness is the property of an item that is difficult to comprehend and differentiate. The principles of vague set theory are used to deal with vagueness. Unlike fuzzy logic which is a special case of vague logic, the vague sets allows to bound the existence of item(s) to an interval. Any item belonging or not will be denoted by its true and false membership. We incorporate vague logic with the classical Apriori algorithm to find more relevant yet vague rules.

VI. CONCLUSION

Vague Association Rule Mining for profit pattern combine the statistic based pattern extraction with value-based decision making to achieve the commercial goals. The work we propose gives an advantage by incorporating vague sets in data mining. The vague sets allow us to consider the vague uncertainty existing in databases and to utilize it to mine such rules that ultimately give better correlation among items. The result calculated was better in comparison to the traditional approach and gives an alternative approach to data mining. The proposed approach not only improves the mining process but also provide the profitable rules in uncertain data. Although a many researches has been carried out in association rule mining but still it requires more attention for defining the notion of profit which would help in improving business strategies and provide some recommender rules.

ACKNOWLEDGMENT

This work is supported by research project under Fast Track Scheme for Young Scientist from DST, New Delhi, India. Scheme 2011-12, No. SR/FTP/ETA-121/ 2011 (SERB), dated 18/12/2012.

REFERENCES

[1] Agrawal Rakesh, Imielinski Tomas, Swami Arun, “Mining Fuzzy Weighted Association Rules” 1993 © SIGMOD ACM

[2] Agrawal Rakesh, Srikant Ramakrishnan, “Fast Algorithms for Mining Association Rules - A Priori”, © 1994

[3] Badhe V, Thakur R.S., Thakur G.S., “A Review on Dealing Uncertainty, imprecision and Vagueness in Association Rule Mining Using Extended and Generalized Fuzzy” IJECS, Vol. 3, issue 7, 2014

[4] Gau Wen-Lung and Buehrer Daniel J., “Vague Sets”, © 1993 IEEE

[5] Han J. and Kamber M., “Data Mining: Concepts and techniques”, Morgan Kaufmann Publishers, Elsevier India, 2001.

[6] Lu, An., Ng,W “Managing merged data by vague functional dependencies”. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. LNCS, vol. 3288, pp. 259–272. Springer, 2004

[7] Lu An and Ng Wilfred ”Maintaining consistency of vague databases using data dependencies” Data and Knowledge Engineering, Volume 68,2009,Pages 622-641.

[8] Lu.A.,Ng.W:Handling Inconsistency of vague relations with functional dependencies. Springer 2007.

[9] Lu An and Ng Wilfred, “Vague Sets or Intuitionistic Fuzzy Sets for Handling Vague Data- Which One Is Better?” 2005 © Springer

[10] Lu An, Ke Yiping, Cheng James, and Ng Wilfred, “Mining Vague Association Rules” 2007 © Springer

[11] Lu An and Ng Wilfred “Mining Hesitation Information by Vague Association Rules”Lecture Notes in Computer Science ,Springer Volume 4801/,2008,pg 39-55.

[12] Pandey Anjana, Pardasani K.R. ,A Model for Mining Course Information using Vague Association Rule, International *Journal of Computer Applications (0975 – 8887), Volume 58– No.20, November 2012*

[13] Pandey Anjana, Pardasani K.R. ,” A Model for Vague Association Rule Mining in Temporal Databases, Journal of Information and Computing Science, Vol. 8, No. 1, 2013, *pp*. 063-074

[14] Pujari A. K., Data Mining Techniques, University Press 2001.

[15] Sandhu, P.S.; Dhaliwal, D.S.; Panda, S.N.; Bisht, A., “An Improvement in Apriori Algorithm Using Profit and Quantity” ICCNT Year: 2010, IEEE conference publication.

[16] Tew. C, Giraud-Carrier C, Tanner K, Burton S. “Behaviour based clustering and analysis of interesting measures for association rule mining” Springer 2013.

[17] Tiwari A., Gupta R.K. and Agrawal D.P. “A survey on Frequent Pattern Mining : Current Status and Challenging issues” Information Technology Journal 9(7) 1278-1293, 2010.

[18] Wang Ke, Zhou Senqiang, and Han Jiawei, Profit Mining: From Patterns to Actions, C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 70–87, 2002.Springer-VerlagBerlin.

Using Moore Dijkstra Algorithm with Multi-Agent System to Find Shortest Path over Network

Basem Alrifai[1], Hind Mousa Al-Hamadeen[2]
Department of Software Engineering, Prince Abdullah Bin Ghazi Faculty of Information Technology, Al-Balqa Applied University,Al-Salt, 19117, Jordan

***Abstract*—finding the shortest path over network is very difficult and it is the target for much research, after many researches get the result in many of algorithm and many a mount based on the performance for these algorithm .Shortest paths problems are familiar problems in computer science and mathematics. In these problems, edge weights may represent distances, costs, or any other real-valued quantity that can be added along a path, and that one may wish to minimize. Thus, edge weights are real numbers and the specific operations used are addition to compute the weight of a path and minimum to select the best path weight.**

In this paper we use the Dijkstra's algorithm with new technique to find the shortest path over network to reduce the time we need to find the best path, in this paper we use node for network with the same value which can be use it to find the shortest path but this depend on the number of transition for every node when the node have high number then the node have the high priority to choose it by using this method we descries the time to find the short path .to make this algorithm more distinguish apply multi-agent system (Automata with multiplicities) to find the short path.

Keywords—multi-agent system; shortest paths problems; Dijkstra Algorithm; Automata with multiplicities

I. Introduction

The most common operation is finding the short path from vertex to another the shortest path from vertex u to vertex v is path with minimum weight we can define the shortest path algorithm as the following [3] Digraph

G = (V, E) where v is vertex and E is edge in the graph the weight function W: E → R Weight of path p = v1 → v2……vk is the sum of the weights of its constituent edges is minimized.

W (p) = ∑ w (vi, vi+1) Formatter will need to create these components, incorporating the applicable criteria that follow.

They are many types for shortest path problem such as single- destination shortest- path problem, single–pair shortest path problem, all pair shortest-path problem.

In this paper we will work on single source shortest path problem from vertex v as the source to all other vertices, we have many algorithm to solve this problem and evaluate the shortest path problem, we will make enhancement to the Dijkstras algorithm by when we have two node have the same value we will added many information to node itself.

In past we choose the node randomly but by using the Dijkstras algorithm enhancement we have decision to choose the node based on the number of transition for the node it self

We added multi agent system as a new technique in addition to the Dijkstras algorithm enhancement we use the automata multiplicities in this method there is addresses for every path

II. Proposed Method

Dijkstra's algorithm, conceived by computer scientist Edsger Dijkstra in 1956 and published in 1959,[1][2] is a graph search algorithm that solves the single-source shortest path problem for a graph with non-negative edge path costs, producing a shortest path tree. This algorithm is often used in routing and as a subroutine in other graph algorithms. The main point for this algorithm it is started at the source vertex s and the tree is T every vertices added to T firstly is start S then the vertices which is closest to S then next closest … etc[11]

A. Dijkstra's Algorithm Enhancement

The enhancement Dijkstra's algorithm is represented when we have more than two edges with the same weight such as (V→U) and (V→W) in this case we choose the vertex which have the maximum transition if (U) has the maximum transition than (W) we use the vertex (U) otherwise we choose (W).

B. Pseudo- code for Dijkstra's Algorithm Enhancement

Dijkstra's Algorithm Enhancement (graph G = (V, E),

```
L = labels for every vertex (weight and path)
For I = 1 to n
    L ($V_I$) = infinity
    L (start) = (0, empty)
   S = {}
While S doesn’t contain end {
U = vertex not in S with minimal L (u)
If there are more than two nodes have the same L (u)
      Choose vertex node u which has the maximum transition
       Add u into S
For all vertices v not in S that u is adjacent to v
       If L (U) + W (U, V) < L (V)
       L (V) = L (U) +W (U, V)
           RETURN L (end)}
```

C. Example for Dijkstra's Algorithm Enhancement

If we have the graph contains 8 vertexes as shown in figure 1 we compute the shortest path by two algorithm Dijkstra's Algorithm and Dijkstra's Algorithm Enhancement

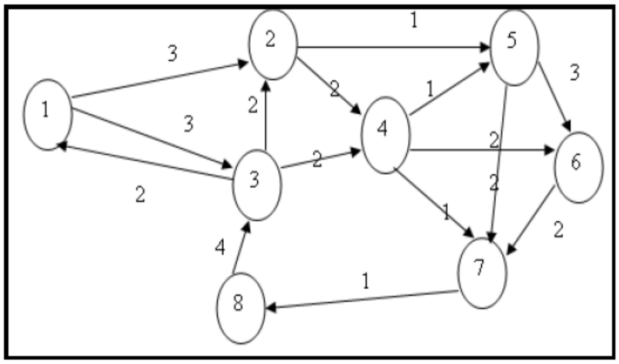


Fig. 1. Example for Dijkstra's Algorithm Enhancement

If we compute the shortest path by using the Dijkstra's Algorithm from source vertex 1 to 8 The path is (1→2→5→7→8) and the shortest path is 7 but if compute the shortest path by using Dijkstra's Algorithm Enhancement we find the path is (1→3→4→7→8) and the shortest path is 7 in this case choose vertex 3 (1 3)because vertex 3 has 3 transition (3→4 ,3→2 , 3→1) whereas vertex 2 has 2 transition (2→ 5,2→ 4) , we have two vertices 2,3 with same value (weight = 3) (1 →3,1 →2)

The difference between two algorithms in Dijkstra's Algorithm the counter for the time is (7) but the Dijkstra's Algorithm Enhancement is (4)

III. Multiagent System and Automata Multiplicities

The multiagent system consists of number of agent, the agent interacts and represent as a user with different goals, and these agents show the ability to cooperate, coordinate and negotiate with each other [1].

An Automaton is advice with permit to assign to every word a coefficient in a smearing and this in an implementable form mainly using matrix computation [8].

We will be able to build effective operation on such automata using of the algebraic structures of the output data [6].

Let k a smearing then an automaton is the data a five up let (Q, A, μ, λ, γ) with :

- Q: the finite set of states
- A: a finite set Alphabet
- μ: A→k Q*Q
- λ € K Q*1 : the set of initial states together with initial values
- γ € k Q*1

Where T = {(q1,a μ(a)q1,q2,q2)} q1,q2 € Q, μ (a) q1,q2# 0[11].

I = {(q € Q :λ (q)#0)}is the set of initial states

F = {(q € Q :λ (q)#0)}is the set of final states

Label (f) =a

Tail (f) =q1

Head (f) =q2

Weight (f) = a

Path c = f1……fm is an element of T

Now we associate the support of T for the weighted graph with edges [6].

Edges (TT) = {(q, a, α, r) € Q* A*K*Q}

This means that every edges (q→r) is superscripted by the pair such that

T(q, a, r) =α #0

The automaton can be observed by means the function it generates this function will be called the behavior of the automaton.

The local behavior of A between two states p;q € Q for the label w € A* is the product of the initial weight λ (p) the total weight of the set of path between p and q with label w and the final weight γ(q) it read [8]

A(p,q) (W) =∑p.q € Qap,q(W)

A. An Agent Modeling Framework Based on Automata Multiplicities

The formalism which is used in our work for the representation of agent behavior produced by perception and action is automata with output the finite inputs alphabet corresponds to the actions set from this output alphabets we build a smearing corresponding to the polynomials over this output alphabet [8]. As we described an automaton with multiplicities over a finite alphabets ∑ and a smearing K is a 5 tuple (∑; Q; I; T; δ) with Q a finite set of states and I, T, δ being mapping such that [8].

I: Q →K
T:Q →K
δ: q* ∑*Q →K

Where I is the set of initial states and T is the set of final states and is δ the transition function Such a structure is useful when transition have outputs to each input word of ∑ is associated an output element of K thus the behavior of an automaton with multiplicities is a series

S = ∑ w€∑*(S │W) W

Where (S │W) is the output elements associated to input word

Example

We have the same graph

Q = {1, 2, 3, 4, 5, 6, 7, 8}
A={a ,b , c}

Now we compute the shortest path by using two algorithms

No.	Counter time for Dijkstra's Algorithm	Counter time for new vision Dijkstra's Algorithm	δ & A

1	9	5	δ={(1,(A,4),2),(1,(b,6),3),(1,(a,4),4),(2,(d,1),5),(3,(b,3),5),(4,(c,2),5),(4,(d,3),6),(5,(b,2),6),(5,(a,4),7),(5,(b,3),8),(6,(d,1),8),(7,(d,2),8),(8,(c,2),9),(9,(b,4),10)} .
			A={a,b,c,d}

2	7	4	δ={(1,(a,3),2),(1,(b,3)3),(2,(1,c),5),(2,(a,2),4),(3,(c,2),5),(4,(d,3),6),(5,(b,2),6),(5,(a,4),7),(5,(b,3),8)(7,(d,2),8),(8,(c,2),9),(9,(b,4(,10)}
			A={a,b,c }

3	6	3	δ={(1,(b,2),3),(2,(b,2),6),(3,(b,4),4),(3,(c,4),5),(3,(b,4),2),(4,(c,1),5),(4,(b,3),7),(4,(c,4),6),(5,(c,3),6),(6,(c,1),7),(7,(c,1),8),(8,(b,4),3)}
			A={a,b,c }

4	11	5	δ={(1,(b,5),2),(1,(c,5),4),(2,(c,2),3),(3,(c,2),5),(4,(b,3),5),(4,(b,2),6),(5,(b,6),6),(5,(b,2),7),(6,(c,3),8),(7,(b,2),8),(8,b,5),10),(9,(c,2)10),(9,(b,3),3),(10,(b,6)11),(10,(b,5),12),(11,(c,2),12),(12,(b,1(,11),}
			A={a, c }

5	10	6	δ={(1,(c,3),2),(1,(a,3),3),(2,(b,1),4),(2,(b,3),5),(3,(a,4),1),(3,(b,2),4),(3,(c,2),6),(4,(a,1),6),(4,(c,3),7),(5,(a,2),7),(6,(b,2),7),(7,(a,2),8),(8,(c,4),10),(8,(b,3),9),(9,(a,2),10),(10,(a,3),1),(11,(b,4),9),}
			A={a,b,c }

Fig. 2. computing the shortest path by using two algorithms

The value of shortest path for this graph from vertex (1) to (8) by using Dijkstra's Algorithm is (7)= (3+1+2+1) in order where the path is (1→2→5→7→8) without any addressing because this algorithm applied without using automata with multiplicities but if we apply new algorithm we find that path is (1→3→4→7→8) with the same value of shortest path (7) =(3+2+1+1) in order but this method we distinguish this path by (bcbc) addressing which is obtained from A ={a, b,c} where edge (1,3) has label (b), edge (7,8) has label (c) edge(4,7) has label b and edge (7,8) has label c so the final addressing for this path (concatenated) is (bcbc)

IV. Results

We used different graph to find shortest path by using two algorithms: Dijkstra's Algorithm and Dijkstra's Algorithm enhancement applied on multiagent system. The performance of our proposed vision of Dijkstra's Algorithm is depending on analysis of algorithm and the criteria which are used for prove performance for our algorithm.

Figure 3 and figure 4 display comparison between two algorithms, we see that the counter time for Dijkstra's Algorithm enhancement applied on multiagent system is less than the Dijkstra's Algorithm.

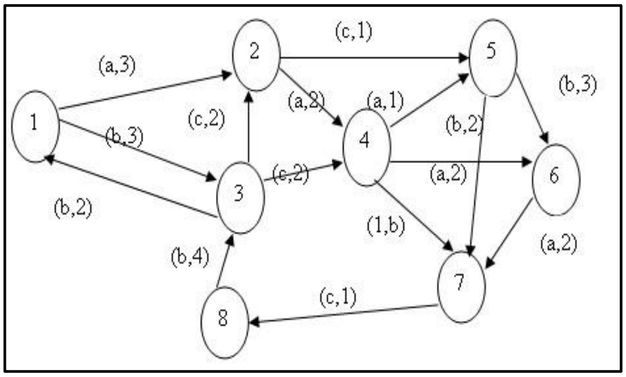


Fig. 3. Comparison between two algorithms

No.	No of vertices	Path in Dj	Path in new vision of Dj	The value of shortest path
1	10	1→2→5→8→9→10	1→4→6→8→9→10	14 from 1 to 10
2	8	1→2→5→7→8	1→3→4→7→8	7 from 1 to 7
3	7	1→3→2→6→7	1→3→4→7	9 from 1 to 7
4	12	1→2→3→8→10→12	1→4→6→8→10→12	20 from 1 to 12
5	10	1→2→4→7→8→10→11	1→3→6→7→8→10→11	16 from 1 to 11

Fig. 4. display path for graph in figure 3

V. Conclusion and Future Work

In this work, enhancement of Dijkstras algorithm for solving single source shortest path problem is discussed.Algorithm tries to solve the problem when have more than two vertices with the same weight and use Multiagent systems techniques.

The properties of this algorithm when we compared with other algorithms can summarized in the following points:

- It is simple and easy to implement over any application
- Uses Q apriority queue ADT

- It is a very Efficient Algorithm to calculate the Shortest Path.
- Uses multi-agent system (simulated by automata with multiplicities) which enhances overall system performance specifically along the dimensions of computational efficiency reliability extensibility maintainability flexibility and reuse

The time required for implementation over any graph by using this new vision of Dijkstras algorithm is less than that by using Dijkstras algorithm .The main shortcoming of this algorithm is this only works if the edges of the graph are nonnegative (Negative weights are not allowed)

In this future we can apply the idea of this method on other algorithm which are used to find shortest path such as(Bellman –Ford algorithm Floyd-Warshall algorithm and Johnsons algorithm) Also we can use other technique for multiagent system with these algorithm which is a specifically representation of automata with multiplicities can be used represent a deterministic agent behavior which is driven by perceptions that induce internal state transitions and can lead to specific action from the agent .(it know transducers) as finite state automata

References

[1] Bordini, R,H bner, and Wooldridge , M2007 Programming Multi-Agent Systems in AgentSpeak using Jason.

[2] Buse, D. Wu, Q2007 .IP Network-based Multi agent system for Industrial Automation . Springer-Verlag London Limitied.

[3] Common ,T Leiserson , c, Rivest, R and Clifford S 2001 .Introduction to Algorithm 2nd Edition,MIT press 2001,PP[.492-508].

[4] G Duchamp, M.Flouret and E langerotte 1999.operation over automata with multiplicities .

[5] Harris,S and Ross, J 2006 Begging algorithm.Wiley publishing ,Inc, Indiana.

[6] Jaff ,L, Bertelle , 2008 shift operation and complex system modinling identification and control Vol.3, No 1,2008.

[7] K culick and J. Kari 1995 Finite state transformations of images.

[8] Wooldridge M,2002 An introduction of multiagent system 2nd John Wiley and Son.

Erp Systems Critical Success Factors

ICT Perspective

Islam K. Sowan
Master of Informatics
Palestine Polytechnic University (PPU)
Hebron, Palestine

Radwan Tahboub
Dept. of Computer Engineering and Sciences
Palestine Polytechnic University (PPU)
Hebron, Palestine

***Abstract*—The Enterprise Resources Planning (ERP) systems are one of the highly complex systems in the information systems field; the implementations of this type of systems need a long time, high cost, and a lot of resources. Many factors affect the successful implementation of ERP system. The critical success factors (CSFs) can be categorized as general, ICT related and software engineering or system life cycle (SLC) related. This paper is a survey paper that identifies ERP systems CSFs in general and software engineering CSFs in specific. Also an agile methodology for ERP systems' implementations will be presented. Many existing ERP systems were surveyed and presented from ICT / software engineering point of view.**

Keywords—ERP; Information and Communication Technology; System Life Cycle; Critical success factors; Agile methodology

I. Introduction

Enterprise resource planning system (ERP) is an information system software that aims to integrate all business processes and functions in a central database; that boosts the management of business resources (finance, production, human resource, materials…etc.) in an effective, efficient, and productive way [1,2,3]. ERP is not a new term; it began in the early of 1960s [4] where the development of ERP starts with Inventory Control (IC) which is an accounting software, then the package is developed to Material Requirements Planning (MRP) during 1970s which gave support for planning and control of production cycle, later MRP was advanced into Manufacturing Resource Planning (MRP II)in 1980s, that aimed to increase the efficiency of manufacturing by technologies integrations for information, the extend of MRP II produce ERP systems[4,5]. Table 1 below depicted the development of ERP systems [6].

TABLE I. Development of ERP SYSTEMS [6]

	Year	Chronology
↑	2009	ERP Cloud
	2000s	Extend ERP
	1990s	ERP
	1980s	MRP II
	1970s	MRP
	1960s	IC

To keep pace with technology, the mobile ERP systems were introduced with the cloud technology [7]. ERP systems have two major advantages [8]: 1.the integrated enterprise show the all business functions and departments; and 2. The centered database eases all business transactions such as the recorded, monitored, and processed. ERP benefits are mentioned in [9] to achieve business benefits by using ERP systems; operational benefits, managerial benefits, strategic benefits, IT infrastructure benefits, and organizational benefits, each main benefits include sub-benefits. Also [7] pointed to eight of mobile ERP systems advantages.

According to [10] there are 67%-90% ERP system failure rate and 35% of ERP implementations are cancelled. Therefore, it is very important to highlight that factors which help and ensure the success of the system; many researches are done to focus on critical success factors (CSFs). The ERP system is a system with a high complex process [11]; critical success factors are used to indicate to the key issues that must be focused by the organizations to ensure a successful system and affects to the implementation process [5]. With a pool of benefits of using ERP systems a lot of studies focused on the impact of using ERP systems on business performance; [12] showed the identification and assessment of the ERP systems' implications and benefits on job performance. In the study the author focus on five factors that can be improved by using ERP systems to achieve outstanding job performance the factors are: 1. Task productivity and innovation, 2. Customer satisfaction, 3. Management control, 4. Interdepartmental communication and cooperation, and 5. Data analysis and conversion.

Many efforts are being made to improve the performance in business and a lot of models are implemented to guarantee gaining a best performance; [13] develop a theoretical model to explain the relationships between each CSF proposed and business performance; the study done for Indians Small Medium Enterprises (SMEs); the authors determine four main CSFs: 1. Approach. 2. Culture. 3. Communication. 4. Support. For each factor have sub-factors. Also five main performance measures are determined in the paper: 1. System quality. 2. Information quality. 3. Organization impact. 4. Workgroup impact. 5. Individual impact. In research [14] they improved a framework include five independent variables financial resource availability, employees perceptions, organizational complexities, regulatory requirements, and having a top management support; and study how these variables affect implementation of an ERP system effectively; and which will have effect on the performance of the firms. The next section will focus on the influence and importance of ICT dimension on this success and in which degree it will help to success. Section three talks about the critical success factors in general those confirm to gain a successful ERP system also mentioned, and the success factors in system life cycle phases in specific; which phases are taken in consideration to support the structure that followed during the implantation journey. Section four proposed agile methodology for ERP systems' implementation is discussed as attempting to find a procedure that is the fastest and requirements changeable to get the user satisfaction and meet the organization requirements. Section number five includes the discussion. Then the conclusion

II. ICT Role for Increasing ERP System Success

A model is assumed in [15] to find the degree of fitness between an organizational dimensions and ERP systems and if they have a power on utilization. In this study they consider the technology is one of the cornerstones to achieve the optimal point and success. They suppose three hypotheses: Technology-Organization Compatibility and Utilization (H1), Technology-Human Compatibility and Utilization (H2), and Human-Organization Compatibility and Utilization (H3). They proved that compatibilities of these dimensions are very important and have a great role on utilization of ERP systems. The adoption of ERP system depends on the process of software selection [6]. The adoption of ICT in the business management increases the efficiency, competitive, and productivity [16].

ICT has a critical role in success of the ERP system; the rate of success of adoption ERP systems in the developing counters is slow and poor when comparing it with others; and this refers to the badness of infrastructure, limited ICT capabilities, and ICT costs. The technical factors are ranked in [10] as one of the critical failure factors in developing counties. One of the most considerable factors in ERP systems implementation is technical complexity [16]. Seeing that risks which related to the ERP project noticed that the technological and implementation issues have the largest share; [17] divide the risks into six major categories, three of them are related to technological and implementation issues: software system design, user involvement and training, and technology planning/integration. Those give us a sight to consider the importance of ICT in developing ERP systems. R.Rajnoha et al. [18] the possible risks during the system life cycle in the ERP system design, implementation, and operation & maintenance phases. All these issue courage us to pour attention to technologies' capabilities, skills, and system implementation methodologies to increase the successful and acceptance of the new ERP systems in the organizations.

In the case study in [1] the National Prawn Company implemented the ERP system two times in 2007 which failed and in 2010 succeeded and met the company requirement; the failure causes that reported are three; two of them refer to software implementation life cycle: 1. No clear vision, goals, and system benefits. 2. Immoderate requirement customization; in the other hand the successful ERP system in 2010 as reported the factors which also related to the technical and software engineering practice, such: more analysis and evaluation steps before the implementation is begin; extra post implementation support "maintenance", more planning issues; training, conversion strategy, and user involvement and stockholders' feedback. Software is one of the ICT components; and the implementation life cycle is a core of success practice of any software the paper will focus on the CSFs in general of the ERP system in the next section and more specific on SLC related success factors. In the table 2 showed the success factors which are related to ICT issues according to the authors.

According to the researches ICT factors are mentioned in the papers as successful factors; which are affected in direct way with ERP systems success; software, communication, network, application development, and project management are some of ICT components which found as a success factors.

III. ERP Critical Success Factors (CSFs)

Although a large number of researches about CSFs, there is no one standard or identical CSFs; each study mentioned a set of different factors than others. This difference occurred within reason of different researches samples and setting [5]. E.Umble et al. [8] proposed nine CSFs, implementation procedure, software selection steps, and a case study for Huck international Inc; and illustrate how the company's ERP system was successful according to the proposed CSFs. In FNah et al. [3] the researchers take 1000 companies' perceptions about CSFs of ERP systems and all factors were evaluated by chief information officers; the results are 11 main CSFs with attached sub-factors and suggest 5 most critical factors: 1. Management support. 2. ERP teamwork and composition. 3. Project management and change management program. 4. Project champion. 5. Change management program and culture. Also the researchers reviewed papers which used the proposed CSFs in their papers. E.Ziemba et al.[19] listed the four groups of critical success factors that are in public administration: 1. Factors related to public procurement procedure. 2. Factors related to government processes management. 3. Factors related to project team competences. 4. Factors related to project management. Each group has underlying factors. Also in [19] showed that CSF for ERP systems according to three different studies; first study: Somers, T. M., & Nelson, K.: "The impact of critical success factors across the stage of ERP Implementation", 2001, they listed 22 CSFs that are related to ERP implementation and did some analysis of these factors according to different phases during the implementation [5]. Second study: Hairul, M., Nasir, N., & Sahibuddin, S. "Critical success factors for software projects: A comparative study", 2011, they divide the CSFs into three groups each one has a number of factors: People related factors, process related factors, and technical related factors. Third study is [20], this study achieves the CSFs that related in implementing lean tools and ERP systems to understand how these CSFs changed over time. E.Ngai et al. [5] represent the literature review of CSFs in complex and systematic way; gather the sub-factors into number of CSFs set according to [3], and reported the CSFs and ERP's performance across countries and regions. R. Addo-Tenkorang [21] showed the literature review of ERP published work in journals between 2005 and 2010.

TABLE II. CSFS THAT RELATED TO ICT ISSUES

	F.Nah et al. [3]	E.Umble et al. [8]	E.Ziemba et al.[19]: Somers and Nelson (2001).	E.Ziemba et al.[19]: Hairul, Nasir, and Sahibuddin (2011).
Software			√	√
Communication	√		√	
Network			√	√
Application development		√	√	
Project management	√	√		√

After a wealth of information about CSFs in general, it turned out to be the largest share of success is for ICT dimensions and software engineering issues and methodologies, hence they can be considered as the backbone of the success of ERP systems. Table 3 mentioned the CSFs that related to SLC issues.

The authors in the table highlighted the factors that are used in software life cycle (SLC) and if you study the system development methodologies you will see and notice the importance of these factors that affect and insure the systems' success in general and ERP systems in special [18]. In other hand some studies that concentrate on the Critical Failure Factors (CFFs); in [10] the aim of the study is to identify CFFs and classify them to avoid ERP systems' failure on Iranian industries fail to help the organization to make an appropriate decision making of ERP implementation and considering these factors would limit the ERP systems failures; they ranked the factors into seven groups: Organizational, Project Management, Human Resources, Managerial, Vendor and Consultant, Processes, and Technical factors.

Project planning and specification it is a phase that related on specification high level of system requirements, business scope, set priorities, complexity, detailed step, duration and work plan, describe the current and new system with risk management plan, data analysis, feasibility studies, Requirements elicitation and analysis, and Requirements validation[22]. In [20], [19: Hairul et al.], and [5] the authors ranked a selection of development processes/methodologies as one of the CSFs. A different methodologies are existing and business needs an importance is various too, therefore the managers should decide to choose the appropriate approach follow, by balancing between technological and business strategies [5]. Project complexity, size and duration mentioned in [19: Hairul et al.] under critical technical-related factors. Understanding of goals, objectives, and business plan: most papers pointed to this factor as a critical one. This factor one of the primary stages that must do [23]. In [5] should be harmonize between the ERP systems missions and business plan, vision, and needs. [8] Clear goals would ensure customer satisfaction, employee will empowered, and facilitate suppliers. Data analysis, conversion, and accuracy; any mistake in data that affects in negative way at the whole final system, the right data entry should have a high priority during the ERP system implementation [8]. Design and development include implementation team, user involvement, and software development. Implementation team is the cornerstone of the ERP system success. All researches mentioned it as a most critical factor the success team refers to the number of members, skills, knowledge, experiences, balancing between technical and business capabilities, and trusted with decision making[3,5,8,19,23]. User involvement increase the acceptance of the new ERP system [23] so it refers to changing management factors too, two types of user involvement reported in [23] the first is involvement in the definition phase and the second in the implementation; hence the user must be involved during all the life cycle phases to increase his satisfaction. Verification and validation are including testing, maintenance, and evaluation. This phase is essential to show the performance of the system[8] and how the operational processes are worked [3], and determination the progress of the system implementation is critical too[5]. Verification and validation generally show the fitting and matching between system and specified requirements in satisfactory manner with users [22]. Education, ERP system is not easy to use by users with limited IT skills [23] and the users of ERP systems are the major cause of success and failure because they will use the system during its life and if they don't know how to use it the rate of failure will increase [8]. In [24] the study done to explain the effects of education on ERP success the study done on 326 firms and the results are all indicated to the importance and positive effects of educations. 1. Increasing the education will increase the ERP success. 2. The increasing of ERP success will increase the organization performance. 3. Statistically education has important effects. Despite of various methodologies for implementing software systems; focus on agile methodology and are highly recommended to be used for ERP development systems because of wide range of flexibility it offers.

IV. PROPOSED AGILE METHODOLOGY FOR ERP SYSTEMS' IMPLEMENTATION

Why Agile? Because the developers are in need to have an ERP system in shortest time and with fastest implementation method; and they need to be flexible with requirements change; with ease of management. Agile can do that. Figure 1 explains the proposed methodology. The above CSFs are considered as the main development phases. The proposed agile method depends on multi development groups' (used two groups), and five phases for each iteration. At the beginning the, small groups were a big one group and they planned the whole project and divided it into increments and iterations. In this stage the group determines the Complexity, project size, duration of the implementation times.

TABLE III. CSFS THAT RELATED TO SLC ISSUES

	F.Nah et al. [3]	E.Umble et al. [8]	Hairul et al. [19]	Somers et al. [19]	O. Alask. et al. [19,20]	E. Ngai et al. [5]	E.Ziemba et al [19]	K.Al-Fawaz [23]
Project planning and Specification(requirements)								
Select a development processes / methodologies.			√		√	√		
Complexity, project size, duration.			√					
Risk management.			√				√	
Understanding of goals, objectives, and business plan.	√	√	√	√	√	√	√	√
Data analysis, conversion, and accuracy.		√	√	√	√	√		
Design and development								
Implementation team.	√	√	√	√	√	√	√	√
User involvement.			√	√		√	√	√
Software development.	√					√		
Verification and validation								
Testing, and troubleshooting.	√					√		
Monitoring and evaluation of performance.	√	√	√			√	√	
Education								
Education and training.		√		√	√	√		√

Risk management is also taking into account. After that the one big group will break into small groups and the teams again gather the requirements from the user for iteration separately understanding the requirement considered the backbone of the success. Each group starts with to implement the independent iteration. But the next group will start when the next finished the first phase; to ensure the ingoing implementation work. And when the group one arrives to the training phase the second constant implemented; thus the implementations will not stopped until the training end. The one iteration takes 1-4 weeks. As knowing the agile can deal with changes and manage it especially requirements' change; each change done by the user the team able to deal with it and modified. Planning the iteration include the requirements that are needed from the iteration. The team will prepare themselves by tools, technologies, and experience. The development team is one of the CSFs that mentioned above during SLC. Time during the life of the phase must be under control any little delay will influence the total project release time. Prototyping is the quick phase to arrange the ERP system requirements in technical from. The user must involve in all software development phases, the regular meetings between the team and users; even the requirements changes remain under control. The next CSF the model support is validation (testing and maintenance), and the additional important phase is the user Education on the developed system. This proposed phase as is essential phase to increase the users' technical abilities of their ERP system [24]. The major goal of ERP systems is to ease the information flow through the organizational then if the users were not educated on the system and trained to solve problems independently the rate of failure will increase. The users are the persons who will use the system and must highlights attention on the way of using it. Also this phase helps the developers to evaluate their work.

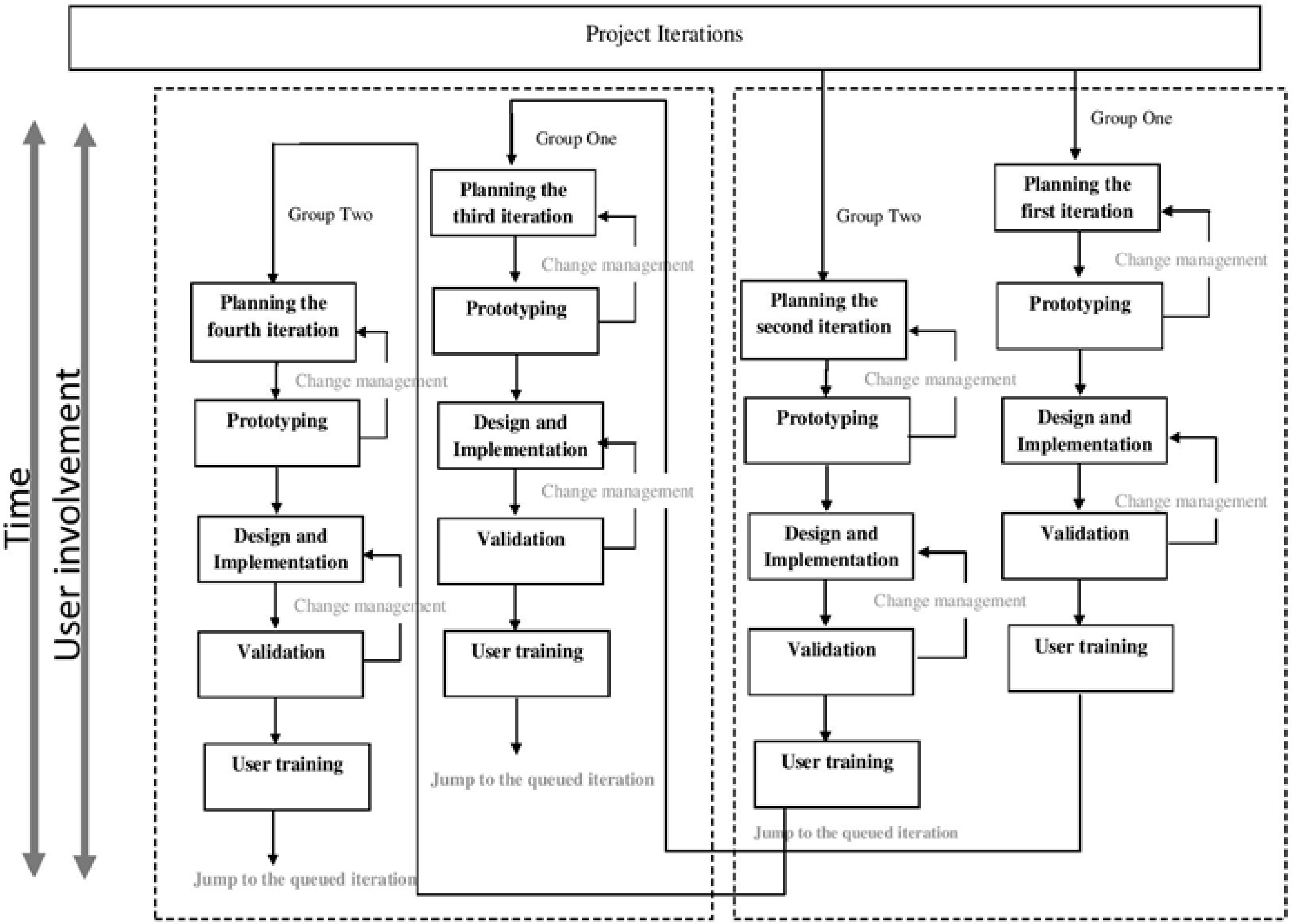


Fig. 1. Proposed Agile Methodology For Erp Systems' Iplementation

V. Discussion

CSFs are the factors that organization must consider to achieve the success during implementing the ERP systems; these types of systems are complex, large and very risky where it may go over budget, and need high managerial capabilities in various levels. Indeed the ERP systems CSFs researches had a plethora of papers dealing with different aspects and characteristics including functional, technical, social, managerial, and implementation features. In this paper, the focusing centered on dimensions that related to ICT success issues and SLC success issues in addition to proposed agile methodology that attempts to improve ERP systems productivity and success. As soon as organizations thought to implement an ERP system they must achieve the best in managerial and technical aspect to gain a successful system with high maximum value.

In [22] social and organizational factors have an impact in requirements for any system, the most important one is culture which has an impact on ERP systems during the life cycle and that impact will also influence the success of the whole ERP, hence, the technical side is not the only effect on the ERP system success, but external social will affect the system as well. Hard to come by distinct ERP system success factors; the ICT and SLC factors are a small side of the pool of factors that related to ERP success.

The high performance of ERP systems is related to the direct relation with the users' performance [24] that lead us to propose the training phase in the modified agile approach to ensure that the user will be able to use the implemented system in the valuable way; if the ERP's system users trained by technical people who implement the system, that will give us a double advantage to produce a well-trained users. In [1] in the mentioned case study the training was the main factor of success in the second successful ERP system for NPC. The importance of the training phase is as important as other phases.

According to this study, the use of ERP systems in educational organizations such universities can be seriously considered; fortune of benefits and facilities that ERP systems offer to both university's employees and students.

VI. Conclusions

Enterprise resource planning system in short (ERP) is information system software that aims to integrate all business process and functions in central database; that increase the management of business resources (finance, production, human resource, materials…etc.) in effective, efficient, and productive way. In this survey paper a comprehensive discussion and review of how different factors affect the success of ERP systems implementations. A summary of these CSFs were presented and compared from different points of view. ICT is also a very important dimension to be considered in ERP systems, this includes software engineering where a customized agile techniques can be used in developing and implementing such systems and their SLC. Also the proposed techniques can be used in implementing ERP systems in educational organizations where huge amount of data and activities need to be managed in an efficient and consistent way.

References

[1] H.Alballaa, A.Al-Mudimigh:"Change Management Strategies for Effective Enterprise Resource Planning Systems: A Case Study of a Saudi Company", international journal of computer applications, 2011.

[2] E.Njihia:"The Effects of Enterprise Resource Planning Systems on Firm's Performance: A Survey of Commercial Banks in Kenya". International journal of business and commerce, 2014.

[3] F.Nah, J.Lau:"ERP implementation: Chief Information Officers' Perception of Critical Success Factors", international journal of human-computer interaction, 2003.

[4] J.Han, R.Liu, B.Swanner, S.Yang:"Executive Summary: Enterprise Resource Planning", 2009.

[5] E.Ngai, C.Law, F.Wat:"Examining the critical success factors in the adoption of enterprise resource palnning". ELSEVIER journal, 2008.

[6] A.Bajahzar, A.Alqahtani, A.Baslem:"A survey study of the Enterprise Resource Planning", international conference on advanced computer science applications and technology, 2013.

[7] Y.Gelogo, H.Kim: "Mobile Integrated Enterprise Resource Planning System Architecture", International Journal of Control and Automation, 2014.

[8] E.Umble, R.Haft, M.Umble:"Enterprise resource planning: Implementation producers and critical success factors", ELSEVIER journal, 2003.

[9] A.Mishra, "Chapter v, Achieving Business Benefits from ERP Systems", 2008.

[10] A. Amid, M.Moalagh, A.Ravasan:"Identification and classification of ERP critical failure factors in Iranian industries", ELSEVIER, 2012.

[11] R.Bhawarkar :"A Framework For The Implementation Of Enterprise Resource Planning (ERP) To Improve The Performance of Business", international journal of research in Advent Technology, 2013.

[12] A. Jalal:"Enterprise Resource Planning: An Empirical Study of Its Impact on Job Performance". International Journal of Business and Information, 2011.

[13] R.Bhawarkar, L.Dhamande:" A FRAMEWORK FOR THE IMPLEMENTATION OF ENTERPRISE RESOURCE PLANNING (ERP) TO IMPROVE THE PERFORMANCE OF BUSINESS". International Journal of Research in Advent Technology, 2013.

[14] E.Njihia.:" The Effects of Enterprise Resource Planning Systems on Firm’s Performance: A Survey of Commercial Banks in Kenya". International Journal of Business and Commerce, 2014.

[15] O. Kerimoglu, N. BaUogluI:" Optimizing the Change Management of Enterprise Resource Planning Systems Implementations". PICMET 2006.

[16] F.A.Goni, A.G.Chofreh, M.Mukhtar, S.Sahran, S.A.Shukor:"Segments and Elements Influenced on ERP System Implementation". Australian Journal of Basic and Applied Sciences, 2012.

[17] M.Sumner:"Risk factors in enterprise-wide/ERP projects".Journal of Information Technology, 2000.

[18] R.Rajnoha, J.Kádárová, A.Sujová, G.Kádár:" Business information systems: research study and methodological proposals for ERP implementation process improvement". 2nd World Conference On Business, Economics And Management, 2014.

[19] E.Ziemba , I.Obłąk: " Critical Success Factors for ERP Systems Implementation in Public Administration ", Interdisciplinary Journal of Information, Knowledge, and Management, 2013.

[20] O.Alaskari, M.Ahmad, N.Dhafr, R.Pinedo-Cuenca:"Critical successful factors (CSFs) for successful implementation of lean tools and ERP systems". Proceedings of the World Congress on Engineering, 2012.

[21] R. Addo-Tenkorang, P. Helo.:" Enterprise Resource Planning (ERP): A Review Literature Report". Proceedings of the World Congress on Engineering and Computer Science, 2011.

[22] I. Sommerville:"Software Engeneering" book; 6th Edition, 2000.

[23] K.Al-Fawaz, Z. Al-Salti, T. Eldabi, :" Critical Success Factors In Erp Implementation: A Review". European and Mediterranean Conference on Information Systems, 2008.

[24] Y. Akça, Ş.Esen, G.Özer:"The Effects of Education on Enterprise Resource Planning Implementation Success and Perceived Organizational Performance". International Business Research.2013

Boosted Decision Trees for Lithiasis Type Identification

Boutalbi Rafika
Computer Science Departement, Badji Mokhtar University
LABGED Laboratory
Annab, Algeria

Chitibi Kheir Eddine, Boutefnouchet
Urology department, CHU Ibn Rochd
Badji Mokhtar University Hospital
Annaba, Algeria

Farah Nadir
Computer Ssience Departement, Badji Mokhtar University
LABGED Laboratory
Annab, Algeria

Tanougast Camel
University of lorraine
LCOMS-ASEC
Metz, France

***Abstract*—Several urologic studies showed that it was important to determine the lithiasis types, in order to limit the recurrence residive risk and the renal function deterioration. The difficult problem posed by urologists for classifying urolithiasis is due to the large number of parameters (components, age, gender, background ...) taking part in the classification, and hence the probable etiology determination. There exist 6 types of urinary lithiasis which are distinguished according to their compositions (chemical components with given proportions), their etiologies and patient profile. This work presents models based on Boosted decision trees results, and which were compared according to their error rates and the runtime. The principal objectives of this work are intended to facilitate the urinary lithiasis classification, to reduce the classification runtime and an epidemiologic interest. The experimental results showed that the method is effective and encouraging for the lithiasis type identification.**

Keywords—urinary lithiasis; classification; Boosting; Decision Trees

I. Introduction

Urinary lithiasis are hard crystals that form in the urinary tract, mostly in the upper urinary tract. From a cooperation with the hospital university center of Annaba (CHU), we obtained a significant related data set. However the major problems of these data resided in their analysis and their interpretations to well define the problem. Physics laboratory of CHU has provided us with data concerning the patients (age, sex, …) and urolithiasis composition. Most collected data are important in determining urinary lithiasis type.

The urolithiasis composition plays a significant role [1] [2] in determining the lithiasis types and their etiologies, which will allow to know the reasons of their occurrence, and help to prescribe a diet or appropriate treatment.

The problem posed in this study was to identify urinary lithiasis type according to their compositions and the patient's profile.

There exist 6 types of urolithiasis which differ according to their morphological and chemical compositions, the six types of urolithiasis are presented in the following figure (Fig. 1). Most studies [3] is based on the four most dominant urinary lithiasis types, namely types 1,2,3 and type 4, because 80% of the urinary lithiasis are part of these four types. In this work the six types of existing urinary lithiases , have been included in the classification, and which correspond respectively to the following etiologies: hyperoxaluria, hypercalciuria, Hyperuricosuria, urinary infections, Cystinuria, Proteinuria.

Each of the six types is composed of the following substances: C1 for type 1. C2 and C1 for the type2. C1, C2 and AU for type 3. C1, C2 and CA for type 4. Cystine and CA for type 5. C1 and Protein for type 6 [3, 4]. However, these six types are not only composed of the quoted components but contain tens other components, with relatively low amounts, which make it possible to effectively distinguish the six types , which is not the case for the components present with large amounts (appendix 1).

In this article, a boosted decision tree system was used to determine the urinary lithiasis types.

This paper is organized as follows. Section II presents the related works. Section III discusses data analysis and data reduction. In Section IV the different methods and tools used were explained. In Section V the results of these methods are presented and compared to other models of learning, according to their classification accuracy, thus etiologies determination. Finally, Section VI concludes the paper.

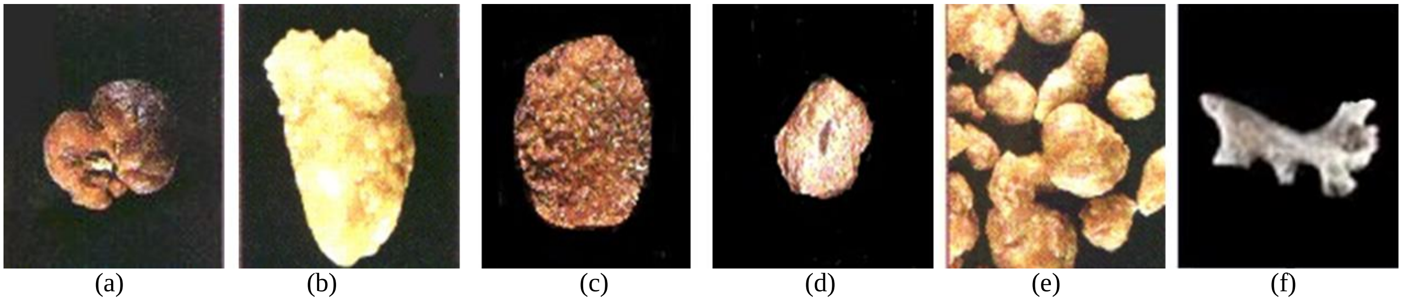


Fig. 1. The six urolithiasis types, (a) Type I, (b) Type II, (c) Type III, (d) Type IV, (e) Type V, (f) Type VI

II. Related Works

Work on various urolithiasis types recognition, have been the subject of some studies, in particular the work performed by Igor Kuzmanovski et al., in their article "Classification of Urinary Calculi using Feed-Forward Neural Network", they carried out, using a neural network, the urolithiasis classification based on lithiasis spectrophotometric analysis. Genetic algorithms have been used to optimize the selection of the most suitable spectral areas in order to improve the classification.

We realized at the beginning of this project, a first work on the classification of urolithiasis, which gave promising results. We presented the results of three classifiers system used: neural network, SVM and a neuro-fuzzy system, and were compared according to their effectiveness.

III. Data Analysis

In this work the data of 528 patients were used, each sample (for each patient) has 23 features which are: age, sex, and twenty-one components (C1, C2, CA, AU0, AU2, WFP, Br, Cystine, URAM, Calcite, Protein, Trg, Mps, Wk, pacc, ocp, urna, Inc, oxypurinol, nexbrit, polysa) (TABLE I).

After having standardized (TABLE II) the 378 patients data, they were divided into two subsets: 378 samples for the training stage and 150 for the validation stage.

TABLE II. Data Codification

Data	Codification
Age and Quantity of components	Integer
Sex	0 for woman 1 for man

Data normalization is an important step especially for classifiers based on distance calculation between two samples like KNN. The normalization ensures that no variable takes too much importance simply because of its measurement unit and it also allows to give equal weight to all the variables. The Normalization of our features was realized using the following formula:

$$fn = \frac{f - f_{min}}{f_{max} - f_{min}}$$

Where fn is a normalized feature value, f is the original feature value, f_{min} is a minimum of feature values and f_{max} is a maximum of feature values.

Several data analysis in particular statistical analysis were performed to better evaluate and interpret data.

Of the 378 cases recorded, there is a ratio man/woman equal to 1.6 i.e. 3 men for 2 women suffering from renal lithiasis (Fig. 2).

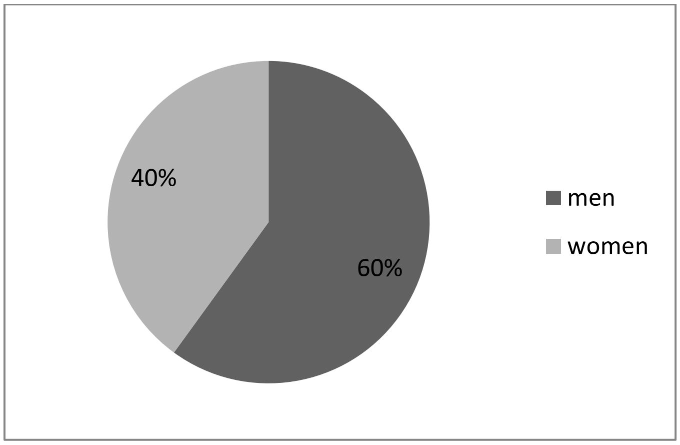


Fig. 2. Proportion of the patients with urolithiasis according to the sex

We found that the average age of calculi appearance in the men population is 47 years and 45ans for the women population(Fig. 3).

Statistical analysis based on urolithiasis type distribution according to the sex, showed that Type 2 mostly dominates in men population while type 4 dominates in women population (Fig. 4).

TABLE I. Data Table

C1	C2	CA	AU0	AU2	PAM	BR	Cys	Prot	Uram	Cal	Trg	Mps	Wk	Pacc	ocp	urna	inc	oxyp	nex	poly
97	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
85	0	5	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
72	20	5	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
7	0	80	0	0	10	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0

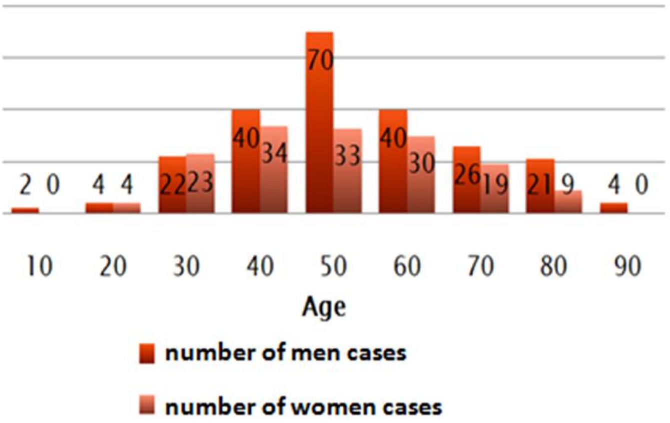


Fig. 3. Number of Cases listed by age, group and sex

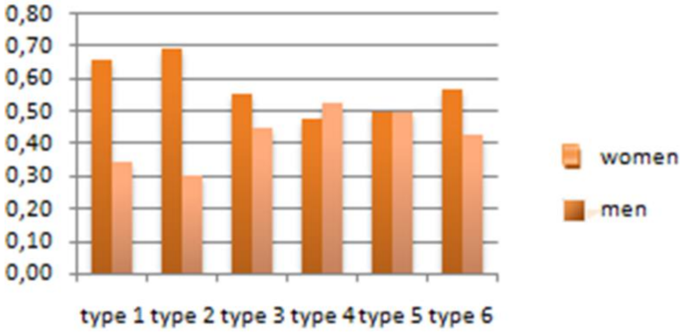


Fig. 4. The urolithiasis type distribution according to the sex

The principal component analysis (PCA) is a method of data analysis family and more generally of multivariate statistics, which involves transforming variables linked together (called "correlated") in new variables decorrelated from each other. These new variables are named "principal components" or "principal axis" . It allows the practitioner to reduce the number of variables and produce the least redundant information[5].

In order to extract as much information as possible and to have a global view on the data, a principal components analysis (PCA) was applied to urolithiasis features to reduce the components number (Fig. 5). The PCA showed that there are correlations between the various components. It is therefore a classification problem with primary variables reduction.

By applying a PCA, we have been able to reduce the number of components from 21 to 11, added to age and sex information, we obtained 13 features for the model (TABLE II).

IV. Methods and Tools

A. *Decision trees*

Decision trees are a type of structures that may deduce a final result from successive decisions. To span a decision tree searching for a solution it is necessary to start from the root. Each node is an atomic decision. Each sub-tree answer allows to move in the one of the child node direction. Gradually, we go down in the tree up to finding a leaf. The leaf represents the answer which the tree gives to the tested sample [6].

The algorithm used to generate decision trees is the C4.5 algorithm, it completely depends on the ID3 algorithm, but has been proposed to overcome the ID3 algorithm limitations.

B. *Boosting*

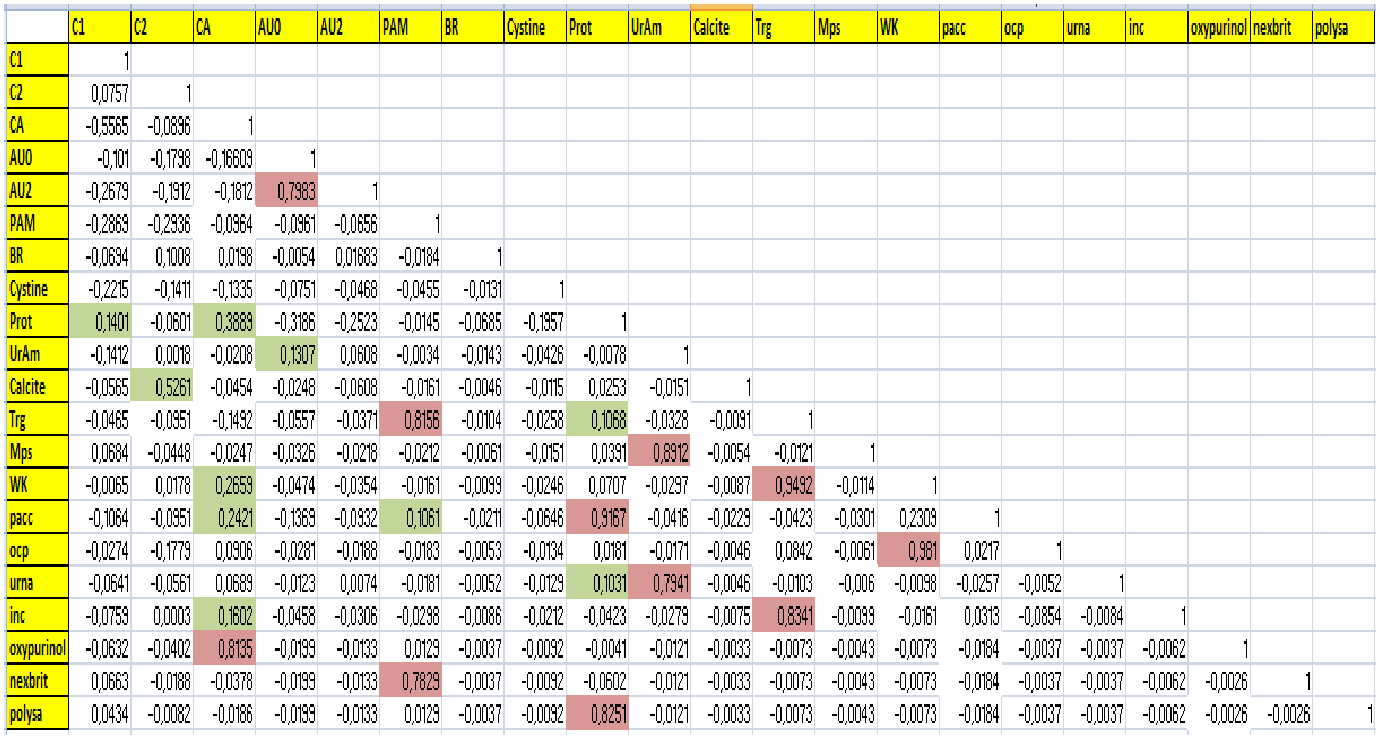
Boosting algorithm [7,8] is a machine learning method, precisely it belongs to meta algorithm family. There are several variations of boosting algorithms, some of them are applied to multiclass problems like AdaBoost.MH [9].

One of the main ideas of AdaBoost, is to set at each steps $1 \leqslant t \leqslant T$, a new prior probability distribution Dt for learning samples based on the algorithm results in the previous step. The weight to "t" step for example (xi, ui) of index i, where xi is sample and ui is a class, is denoted pt (i). Initially, all examples have the same weight, then at each step the weights of misclassified examples are increased, forcing the learner to focus on the difficult examples of the training sample.

Many classification studies [8, 9] showed only the Boosting algorithm effectiveness on simple decision rules.

In order to experimentally select a best decision trees for the boosting algorithm, many decision trees have been generated separately. The boosting algorithm was performed on these decision trees.

C. Proposed method



	C1	C2	CA	AU0	AU2	PAM	BR	Cystine	Prot	UrAm	Calcite	Trg	Mps	WK	pacc	ocp	urna	inc	oxypurinol	nexbrit	polysa
C1	1																				
C2	0,0757	1																			
CA	-0,5565	-0,0896	1																		
AU0	-0,101	-0,1798	-0,16609	1																	
AU2	-0,2679	-0,1912	-0,1812	0,7983	1																
PAM	-0,2869	-0,2936	-0,0964	-0,0961	-0,0656	1															
BR	-0,0694	0,1008	0,0198	-0,0054	0,01683	-0,0184	1														
Cystine	-0,2215	-0,1411	-0,1335	-0,0751	-0,0468	-0,0455	-0,0131	1													
Prot	0,1401	-0,0601	0,3889	-0,3186	-0,2523	-0,0145	-0,0685	-0,1957	1												
UrAm	-0,1412	0,0018	-0,0208	0,1307	0,0608	-0,0034	-0,0143	-0,0426	-0,0078	1											
Calcite	-0,0565	0,5261	-0,0454	-0,0248	-0,0608	-0,0161	-0,0046	-0,0115	0,0253	-0,0151	1										
Trg	-0,0465	-0,0951	-0,1492	-0,0557	-0,0371	0,8156	-0,0104	-0,0258	0,1068	-0,0328	-0,0091	1									
Mps	0,0684	-0,0448	-0,0247	-0,0326	-0,0218	-0,0212	-0,0061	-0,0151	0,0391	0,8912	-0,0054	-0,0121	1								
WK	-0,0065	0,0178	0,2659	-0,0474	-0,0354	-0,0161	-0,0099	-0,0246	0,0707	-0,0297	-0,0087	0,9492	-0,0114	1							
pacc	-0,1064	-0,0951	0,2421	-0,1369	-0,0932	0,1061	-0,0211	-0,0646	0,9167	-0,0416	-0,0229	-0,0423	-0,0301	0,2309	1						
ocp	-0,0274	-0,1779	0,0906	-0,0281	-0,0188	-0,0183	-0,0053	-0,0134	0,0181	-0,0171	-0,0046	0,0842	-0,0061	0,981	0,0217	1					
urna	-0,0641	-0,0561	0,0689	-0,0123	0,0074	-0,0181	-0,0052	-0,0129	0,1031	0,7941	-0,0046	-0,0103	-0,006	-0,0098	-0,0257	-0,0052	1				
inc	-0,0759	0,0003	0,1602	-0,0458	-0,0306	-0,0298	-0,0086	-0,0212	-0,0423	-0,0279	-0,0075	0,8341	-0,0099	-0,0161	0,0313	-0,0854	-0,0084	1			
oxypurinol	-0,0632	-0,0402	0,8135	-0,0199	-0,0133	0,0129	-0,0037	-0,0092	-0,0041	-0,0121	-0,0033	-0,0073	-0,0043	-0,0073	-0,0184	-0,0037	-0,0037	-0,0062	1		
nexbrit	0,0663	-0,0188	-0,0378	-0,0199	-0,0133	0,7829	-0,0037	-0,0092	-0,0602	-0,0121	-0,0033	-0,0073	-0,0043	-0,0073	-0,0184	-0,0037	-0,0037	-0,0062	-0,0026	1	
polysa	0,0434	-0,0082	-0,0186	-0,0199	-0,0133	0,0129	-0,0037	-0,0092	0,8251	-0,0121	-0,0033	-0,0073	-0,0043	-0,0073	-0,0184	-0,0037	-0,0037	-0,0062	-0,0026	-0,0026	1

Fig. 5. correlation matrix

TABLE III. FINAL DATA

Age	Sex	C1	C2	CA	AU0	AU2	PAM	Cystine	Prot	Uram	pacc	Other
41	1	0	85	15	0	0	3	0	0	0	0	0
79	1	5	0	0	70	25	0	0	0	0	0	0
50	0	0	40	55	0	0	0	0	5	0	0	0
22	0	78	15	3	0	0	0	0	2	0	2	0
34	1	7	3	40	0	0	15	0	12	0	20	3

The idea of this work consists in implementing the algorithm of boosting on decision trees. Two stages are required, first decision trees creation and then boosting algorithm application. Decision trees were directly generated from data files randomly created. However, two important factors must be taken into consideration:

- Tree depth
- Number of trees

We decided to experiment the boosting on small decision trees, constrained by their depth. Deep enough to separate data and make a decision, but not too deep to maintain general rules and avoid over-learning.

The number of trees used in boosting must be fixed, not too large for not to slow down training step, and not too small for boosting algorithm powerfulness.

V. RESULTS AND DISCUSSIONS

The used model, boosting of decision tree, generates different results according to the selected parameters. For evaluated system, two parameters must be fixed: the tree depth and the number of trees. The results are presented in terms of training rate error and runtime.

A. Evaluation according to trees depth variation

The decision tree depth used in boosting algorithm varies from 3 to 5.

TABLE IV shows the obtained results of our model while varying the depth for a boosting with 15 decision trees.

The results in TABLE IV shows that when using low trees depth (depth 3) i.e. with the simpler rules, we obtain better performances than trees with large depth (depth 5), however the runtime for small tree depth is almost doubled; 500ms for the tree with depth 3 and 249 ms for the tree with depth 5.

TABLE IV. RESULTS ACCORDING TO DECISION TREE DEPTHS

Depth	Error rate	Time(ms)
Depth 5	9%	249
Depth 4	4,5%	430
Depth 3	1,59%	500

B. *Evaluation according to the number of trees variation*

The number of decision trees used in the Boosting algorithm takes on the three following values: 10, 15 and 20 trees.

TABLE V illustrates the results obtained by our model while varying the number of trees for Boosting under a fixed depth equal to 4. It is shown that with a greater number, learning gives better results and therefore a lower error rate.

TABLE V. RESULTS ACCORDING TO NUMBER OF DECISION TREES

Number of decision trees	Error rate	Time(ms)
10 trees	8%	249
15 trees	4,5%	374
20 trees	2%	455

Compared to our parameters, trees depth (three possible values) and the number of trees (three possible values), you can have 9 different combinations and therefore 9 different systems according to their error rate and their runtime (Fig. 5).

Fig. 6 presents the error rate of each of the nine system combination. The system that gives the best result is the one with 15 decision trees and depth equal to 3.

The most powerful model, a compromise between execution time minimization and error rate, is the one with 15 decision trees and depth equal to 3. The details of this model and its confusion matrix are presented in Fig. 7. It happens to reach a classification accuracy equal to 98.41% , with a correct classification rate of 100% for types 3 and 5 and 99% for types 1 and 2. The execution time is 500ms.

In Fig. 8, the blue curve represents the error rate of the learning stage and the red curve represents the error rate of the validation step. The validation error is almost equivalent to the learning error, our system is efficient, with an error rate equal to 1.59% for the training step and a rate error equal to 1.35% for the validation step. The iteration number is approximately equal to 400 iterations for the two steps.

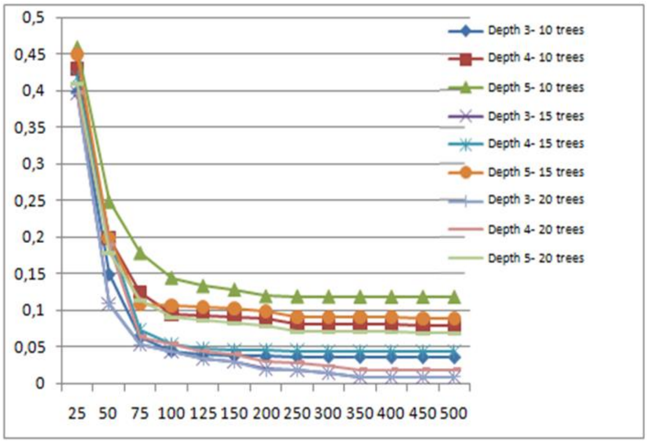


Fig. 6. Curves of results according to depth and number of trees used for Boosting based on error rate and iterations number

VI. CONCLUSION

In conclusion, this work has allowed us to achieve our objectives, namely the effective classification of urolithiasis. The boosting model proposed using 15 decision trees with a depth equal to 3 is the best one for this classification problem. Its accuracy is 98.41% for the urolithiasis classification. He correctly classified 372 cases of 378 cases.

This model with a validation error equal to 1.35%, can be considered as a promising model for the identification of urinary tract stones and determination of étiologies.

REFERENCE

[1] A. Hesse, M. Gergeleit, P. Schüller and K. Möller, 'Analysis of Urinary Stones by Computerized Infrared Spectroscopy', *Clinical Chemistry and Laboratory Medicine*, vol. 27, no. 9, 1989.

[2] V. M, d. JC and G. HM, 'Infrared analysis of urinary calculi by a single reflection accessory and a neural network interpretation algorithm.', *Clinical chemistry*, vol. 47, no. 7, pp. 1287-1296, 2000.

[3] J. Guerra-López, J. Güida and C. Della Védova, 'Infrared and Raman studies on renal stones: the use of second derivative infrared spectra', *Urological Research*, vol. 38, no. 5, pp. 383-390, 2010.

[4] I. Kuzmanovski, M. Trpkovska and B. optrajanov, Maked. Med. Pregled, 1999, 53, 251–255.

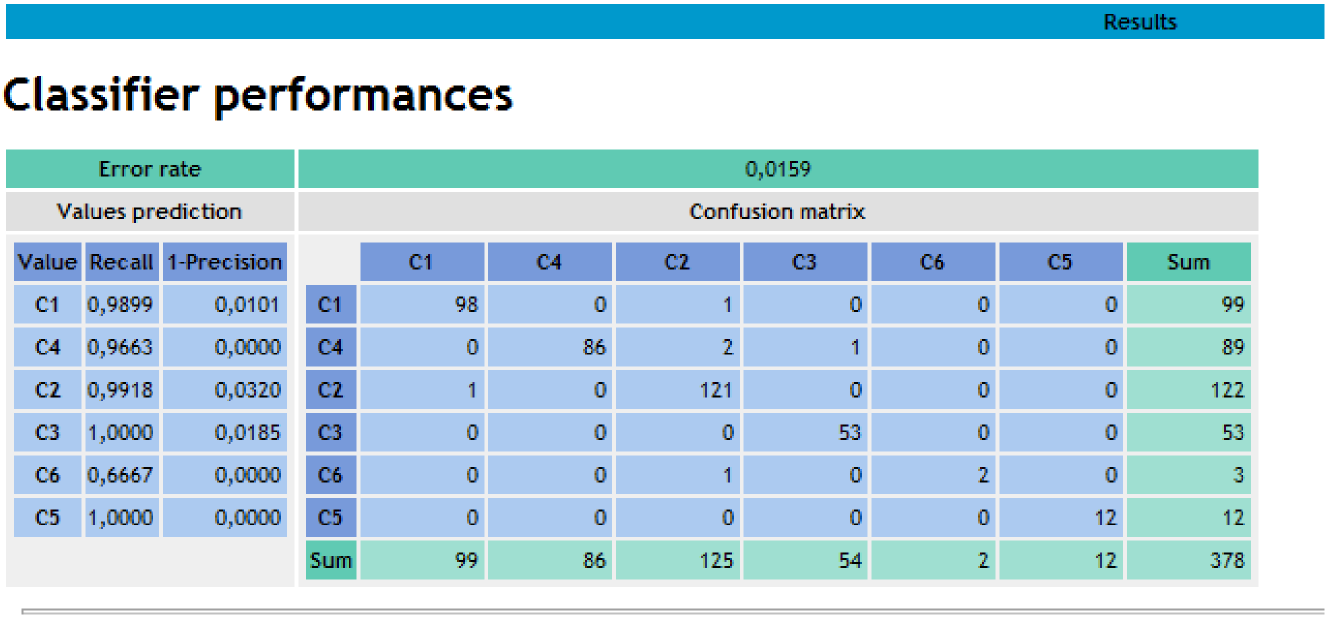
[5] H. Abdi and L. Williams, 'Principal component analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 2010.

[6] J. Quinlan, *C4.5*. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[7] Y.FreundandR.Schapire, "Experimentswithanewboostingalgorithm," Proc. 13th Int. Conf. Mach. Learn.,San Mateo, CA: Morgan Kaufmann, 1996,pp.148–156.

[8] J. Quinlan, 'Bagging, Boosting, and C4.5', *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 2015.

[9] Y. Freund and R. E. Schapire, "A short introduction to boosting", *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp.771 -780 1999.



Results

Classifier performances

Error rate	0,0159
Values prediction	Confusion matrix

Value	Recall	1-Precision
C1	0,9899	0,0101
C4	0,9663	0,0000
C2	0,9918	0,0320
C3	1,0000	0,0185
C6	0,6667	0,0000
C5	1,0000	0,0000

	C1	C4	C2	C3	C6	C5	Sum
C1	98	0	1	0	0	0	99
C4	0	86	2	1	0	0	89
C2	1	0	121	0	0	0	122
C3	0	0	0	53	0	0	53
C6	0	0	1	0	2	0	3
C5	0	0	0	0	0	12	12
Sum	99	86	125	54	2	12	378

Fig. 7. Boosting model results with 15 decision trees and depth equal to 3

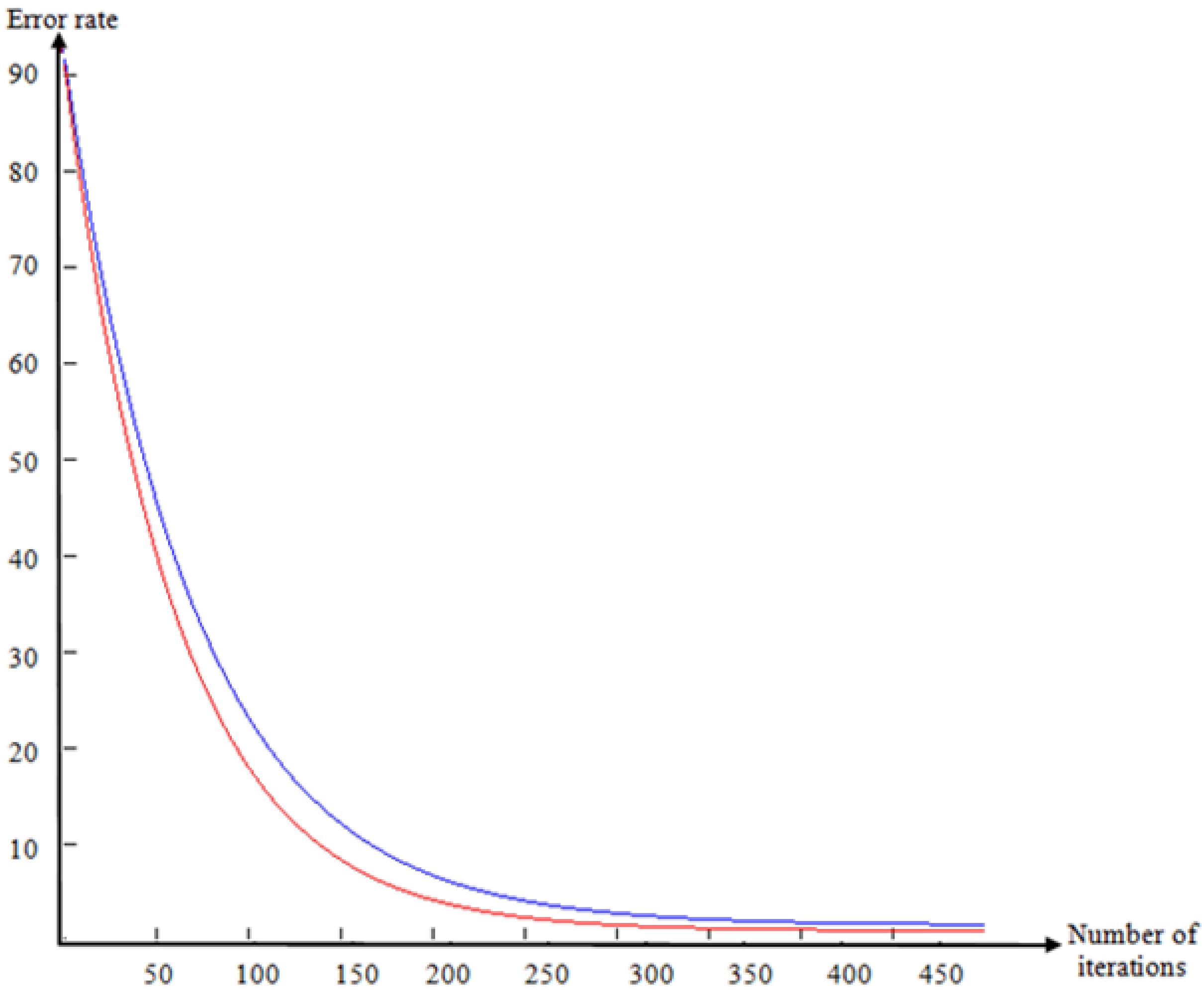


Fig. 8. Error rate Curves of training and validation steps

Meteosat Images Encryption based on AES and RSA Algorithms

Meteosat Image Encryption

[1]Boukhatem Mohammed Belkaid
Laboratoire d'Analyse et
Modélisation des Phénomènes
Aléatoires, UMMTO, BP 17 RP,
15000,
Tizi-Ouzou, Algérie

[2]Lahdir Mourad
Laboratoire d'Analyse et
Modélisation des Phénomènes
Aléatoires, UMMTO, BP 17 RP,
15000,
Tizi-Ouzou, Algérie

[3]Cherifi Mehdi
Laboratoire d'Analyse et
Modélisation des Phénomènes
Aléatoires, UMMTO, BP 17 RP,
15000,
Tizi-Ouzou, Algérie

***Abstract*—Satellite image Security is playing a vital role in the field of communication system and Internet. This work is interested in securing transmission of Meteosat images on the Internet, in public or local networks. To enhance the security of Meteosat transmission in network communication, a hybrid encryption algorithm based on Advanced Encryption Standard (AES) and Rivest Shamir Adleman (RSA) algorithms is proposed. AES algorithm is used for data transmission because of its higher efficiency in block encryption and RSA algorithm is used for the encryption of the key of the AES because of its management advantages in key cipher. Our encryption system generates a unique password every new session of encryption. Cryptanalysis and various experiments have been carried out and the results were reported in this paper, which demonstrate the feasibility and flexibility of the proposed scheme.**

***Keywords*—AES; RSA; MSG; satellite; encryption; keys**

I. Introduction

The amount of satellite image has increased rapidly on the Internet, in public or local networks. Meteosat image security becomes increasingly important for many applications, e.g., confidential transmission, multispectral imaging for providing electronic images of clouds, land and sea surfaces, analysis of air masses to monitor the thermodynamic state in the lower part of the atmosphere and environment data collection and relay transmitted by automatic platforms (marine beacons, land and airborne ...)[1]. The unlawful, unofficial, and unauthorized access and illegal use of Meteosat imagery increases the importance of information security to keep the critical and confidential imagery and transmission process secure, dependable, trustworthy, and reliable. Cryptography is the most widely accepted information security technique employed to make the Meteosat image transmission processes reliable and secure from unauthorized access and illegal use [2-3]. Cryptographic techniques can be divided into symmetric (with a secret key) and asymmetric encryption (with private and public keys). In symmetric cryptosystems, the same key is used for the encryption or decryption and this key need to be secure and must be shared between the transmitter and the receiver. These cryptosystems are very fast and easy to use. Many image encryption algorithms have been developed in last year’s. Among them, we find, the public symmetric AES algorithm, which has proven its robustness against different types of attacks nowadays [4-9], the asymmetric RSA [10-12] algorithm and the IDEA algorithm. Using these algorithms allow separately kind of luxurious ensure confidentiality. For this reason, a hybrid cryptosystem based on both AES and RSA is proposed. The Advanced Encryption Standard (AES) and the Rivest Shamir Adleman (RSA) algorithms are the two popular encryption algorithms that vouch confidentiality, integrity and authenticity over an insecure communication network and Internet.AES algorithm which contain iterative rounds. AES algorithm support several cipher modes of operation such as ECB (Electronic Code Book), CBC (Cipher Block chaining), OFB (Output Feedback), CFB (Cipher Feedback) and CTR (Counter) [13-15]. In our system, privacy is ensured by AES algorithm using five modes of operation and the RSA algorithm is used to transmit the keys. The cryptosystem also check the integrity of images using a simple process based on correlation between the pixels of Meteosat images. The rest of the paper is organized as follow. Section 2 discusses the proposed hybrid cryptosystem scheme. Section 3 and 4 shows some numerical results. Finally, section 5 concludes the paper.

II. the Cryptosystem Proposed

In this work a communication system based on AES and RSA algorithms is realized. The global scheme of the proposed system for private communications is shown in Fig.1. Note that the transmission channel is a public one. Consequently, any hacker has a free access to information passing through the channel which is considered perfect in our works. The cryptosystem is designed to protect MSG images transmitted over the channel of transmission against any attack.

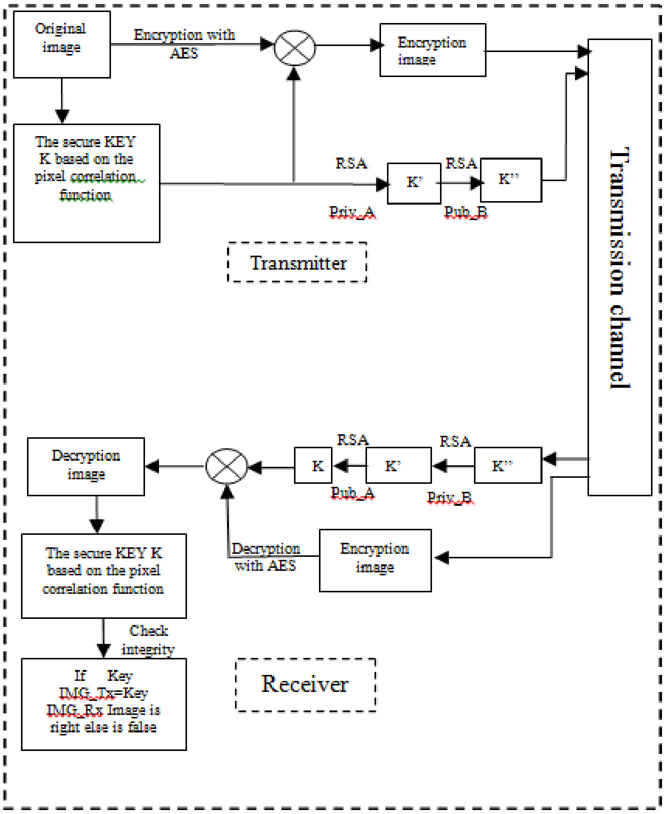


Fig. 1. Transmission chain based on AES and RSA

A. Transmission block

1) AES algorithm structure

The Advanced Encryption Standard (AES) is a specification for the encryption of electronic data established by the U.S. National Institute of Standards and Technology (NIST) in 2001. The figure 2 shows the AES cipher in detail,

Indicating the sequence of transformations in each round and showing the corresponding decryption function. Four different stages are used, one of permutation: ShiftRows, and three of substitution: (1) Substitute bytes, (2) MixColumns, (3) AddRoundKey, and is fast in both software and hardware.

AES is a variant of Rijndael which has a fixed block size of 128 bits, and a key size of 128, 192, or 256 bits. By contrast, the Rijndael specification per se is specified with block and key sizes that may be any multiple of 32 bits, both with a minimum of 128 and a maximum of 256 bits.

2) Cipher operation block

A mode of operation is a technique for adapting the algorithm for an application, such as applying a block cipher to a sequence of data blocks or a data stream. Five modes of operation have been defined by NIST (SP 800-38A) are used. A mode of operation is a technique for enhancing the effect of a cryptographic algorithm or adapting the algorithm for an application, such as applying a block cipher to a sequence of data blocks or a data stream. The five modes are intended to cover a wide variety of applications of encryption for which a block cipher could be used.

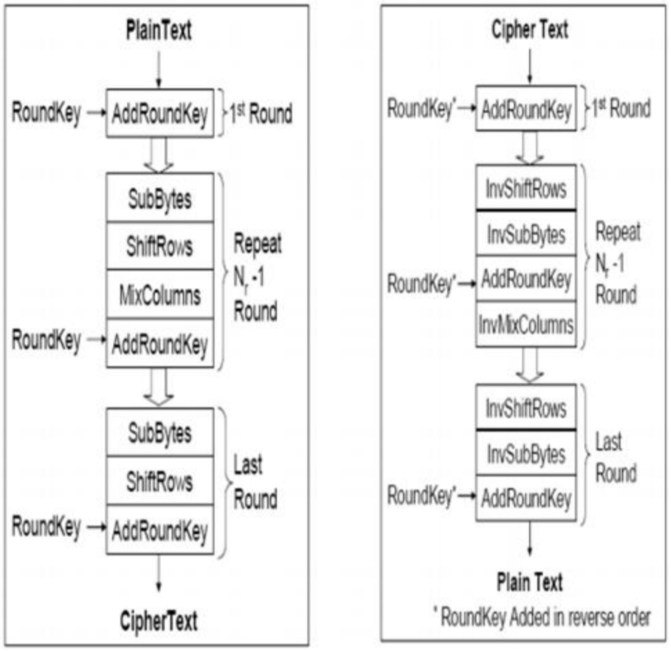


Fig. 2. AES encryption and decryption algorithm

These modes are intended for use with any symmetric block cipher, including triple Data Encryption Standard (DES) [16-17] and AES.

- Electronic codebook (ECB)

Encryption:

$$C_j = E(K, P_j\ ) \qquad j = 1, \ldots.., N \qquad (1)$$

Decryption:

$$P_j = D(K, C_j\ )\ \ j = 1, \ldots.., N \qquad (2)$$

- Cipher block chaining (CBC)

Encryption:

$$C_j = E\left(K, \left[C_{j-1} \oplus P_j\right]\right) \qquad (3)$$

Decryption:

$$\left.\begin{array}{c} D(K, C_j) = D\left(K, E\left(K, \left[C_{j-1} \oplus P_j\right]\right)\right) \\ D(K, C_j) = C_{j-1} \oplus P_j \\ (4) \\ C_{j-1} \oplus\ D(K, C_j) = C_{j-1} \oplus C_{j-1} \oplus P_j = P_j \end{array}\right\}$$

- Cipher feedback (CFB)

Encryption:

$$\left.\begin{array}{l} I_1 = IV \\ I_j = LSB_{b-s}(I_{j-1})//C_{j-1}\ j = 2, \ldots., N \qquad (5) \end{array}\right\}$$

$$I_j = E(K, I_j)\ \ j = 1, \ldots.., N$$

$$C_j = P_j \oplus MSB_s(O_j) \qquad j = 1, \ldots., N$$

Decryption:

$$\left.\begin{aligned} & I_1 = IV \\ & I_j = LSB_{b-s}(I_{j-1})//C_{j-1}\ j = 2, \ldots., N \quad (6) \\ & O_j = E(K, I_j)\ j = 1, \ldots., N \end{aligned}\right\}$$

$$P_j = C_j \oplus MSB_s(O_j) \qquad j = 1, \ldots., N$$

- Output feedback (OFB)

Encryption:

$$C_j = P_j \oplus E(K, [C_{j-1} \oplus P_{j-1}]) \qquad (7)$$

Decryption:

$$C_j = C_j \oplus E(K, [C_{j-1} \oplus P_{j-1}]) \qquad (8)$$

- Counter (CTR)

Encryption:

$$I_1 = Nonce$$

$$\left.\begin{aligned} & I_j = O_{j-1} \qquad j = 2, \ldots, N \\ & O_j = C_j \oplus E(K, I_j)\ \ j = 1, \ldots, N \quad (9) \\ & O_j = E(K, I_j) \qquad j = 1, \ldots, N \end{aligned}\right\}$$

$$C_j = P_j \oplus E(K, I_j) \qquad j = 1, \ldots, N-1$$

Decryption:

$$\left.\begin{aligned} & I_1 = Nonce \\ & I_j = LSB_{b-s}(I_{j-1})//C_{j-1}\ j = 2, \ldots., N\} \\ & O_j = C_j \oplus E(K, I_j) \qquad j = 1, \ldots, N \qquad (10) \\ & P_j = C_j \oplus O_j \qquad j = 1, \ldots, N-1 \end{aligned}\right\}$$

$$P_N^* = C_N^* \oplus MSB_u(O_N)$$

3) *RSA asymmetric algorithm*

The RSA algorithm was publicly described in 1977 by Ron Rivest, Adi Shamir, and Leonard Adleman. The RSA algorithm is the most popular and proven asymmetric key cryptographic algorithm. The RSA algorithm is based on the mathematical fact that it is easy to find and multiply large prime numbers together, but it is extremely difficult to factor their product. The private and public keys in the RSA are based on very large (made up of 100 or more digits) prime numbers [10-12]. In such a cryptosystem, the encryption key is public and differs from the decryption key which is kept secret. In RSA, this asymmetry is based on the practical difficulty of factoring the product of two large prime numbers, the factoring problem. To transmit the key K, the transmitter can encrypt this key using the RSA asymmetric algorithm .The transmitter have the public and private key, $Pub_E(b_x,n_x)$, $Priv_E(u_x,n_x)$, and the receiver have the public and private key $Pub_R(b_y,n_y)$, $Priv_R(u_y,n_y)$.

The transmitter signs the key K with the RSA algorithm using the private key of the sender $priv_E$ to obtain a signed key K' such that:

$$K' = K^{Ux} mod(n_x) \qquad (11)$$

The key K' is encrypted for the second time using the RSA public key Pub receiver to generate the key K":

$$K'' = K'^{by} mod(n_y) \qquad (12)$$

B. *Reception block*

Inverse functions are used to reconstruct the same sent image. Here the function of correlation between adjacent pixels is used to verify integrity. The cryptosystem developed can detect in the reception if a change affects the image in the transmission channel, using the correlation function in the block verification of integrity.

III. Numerical Results

In this study, a Meteosat image database is used. Meteosat images used recorded by the Meteosat Second Generation (MSG) on twelve visible and infrared channels are provided by the meteorological station of the National Meteorology Office (ONM) Dar El Beida, Algeria. Figure 3 shows all MSG images used for our various tests.

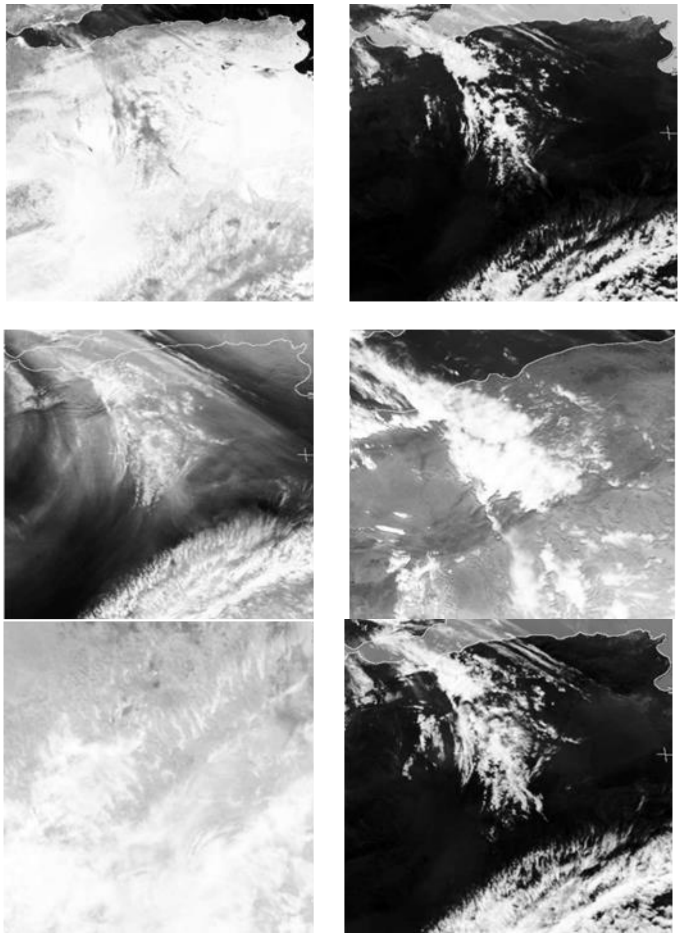


Fig. 3. MSG images in different channels

An ordinary computed Meteosat image, as shown in Figure 4, having a size of 262 144 bytes and a resolution of 512 × 512, is used for the experiments and analysis.

The encrypted and decrypted images aregiven in Figures 5 and 6, respectively, to prove the robustness and quality of the encryption results. The encrypted Meteosat image is totally scrambled and highly secure from unauthorized access and illegal use. The decrypted Meteosat image is the same as the original image, with no changes and/or alterations.

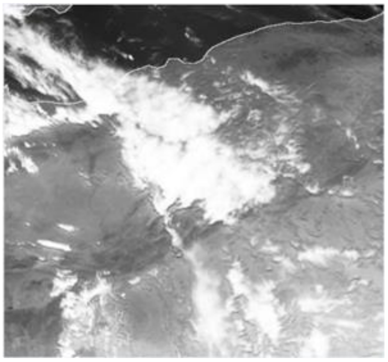


Fig. 4. Original image

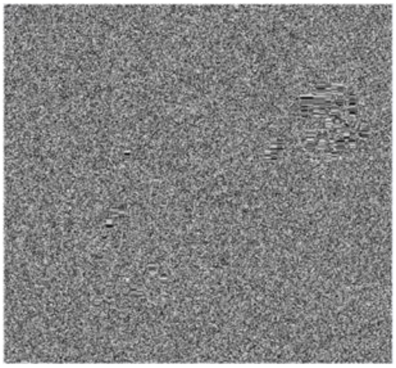


Fig. 5. ECB image encrypted

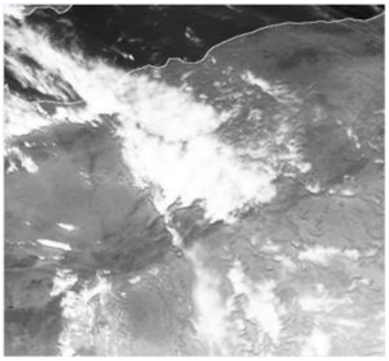


Fig. 6. Decrypted image

IV. Security Analysis

The security of the above-described encryption scheme is now analyzed by studying various tests: histogram analysis, correlation coefficients analysis and key space analysis.

A. *Histogram Analysis*

Figs. 7, 8 and 9 show histograms of an original image and encrypted images for two modes of operation ECB and OFB. The experiment results show that the histogram of the encrypted Meteosat images is fairly uniform and different from the original image.

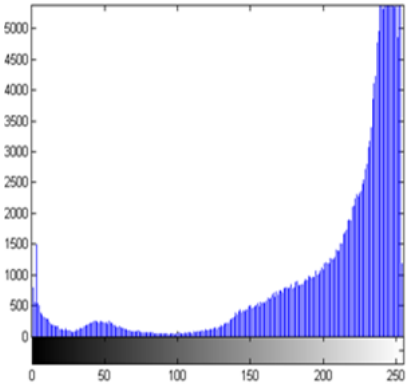


Fig. 7. Histogram of original image

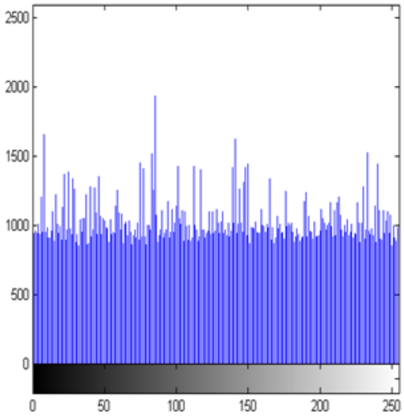


Fig. 8. Histogram of ECB image encrypted

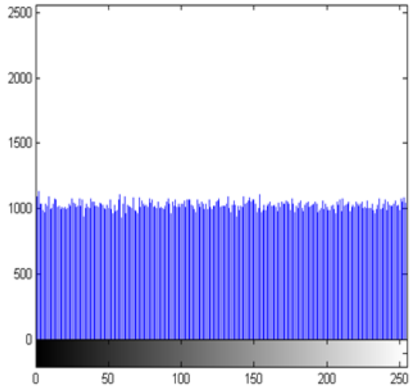


Fig. 9. Histogram of OFB image encrypted

B. *Correlationciefficients analysis*

Figure 10 shows the correlation coefficients for the encrypted Meteosat images for the five modes. It is clear from computed experimental results of these figures that there is negligible correlation between these images. We note that the performance of CBC and CTR modes because they have a lower correlation coefficient. ECB mode has the highest coefficient.

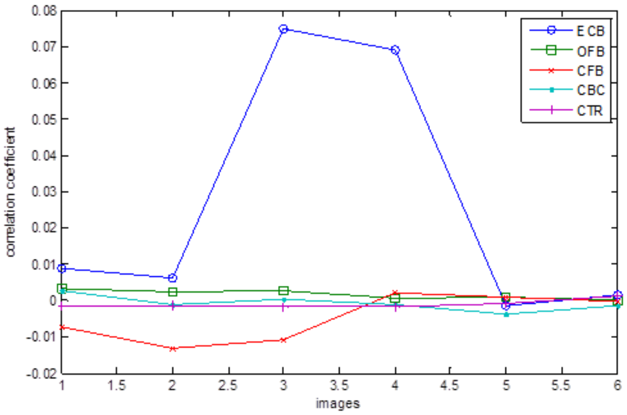


Fig. 10. Correlation coefficients of encrypted images

C. Keysensitivity

Security keys are extremely important to an image encryption algorithm for ensuring the security of protected images in against the differential and brute force attacks. Generally speaking, the security of an image encryption algorithm depends on its security key design. An encryption algorithm should contain a sufficiently large key space and should be strongly sensitive to the change of security keys. Here, the sensitivity tests performance of the encryption and decryption processes as shown in Fig. 11.

As can be seen in Fig. 11 that the five modes have low correlation, except the ECB mode, which the pixels are higher correlated than the others modes.

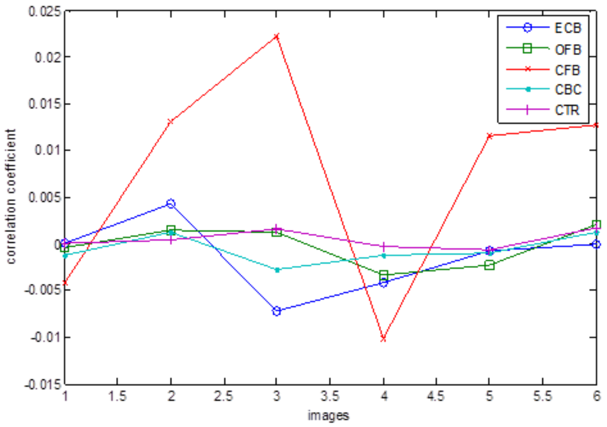


Fig. 11. The sensitivity of Key

D. Integrity Check

For this test, the emission and reception footprint are calculated for the six Meteosat images in CTR mode. The obtained results show in the Table I.

From Table I, the problem of integrity is checked when the image change in the transmission channel because the image of the cryptographic decrypted footprint is different from that of the original images.

TABLE I. INTEGRITY RESULTANT

MSG Images	Footprint emission	Footprint reception
1	16964897393897	14203735824732
2	87151338539745	96443232263919
3	19947467820021	98621115981365
4	19026974966893	16175845459547
5	22994972443810	12356287082727
6	03660222107272	87780179090053

V. CONCLUSION

In this paper, to overcome security, performance, privacy and reliability issues of satellite MSG imagery, a new cryptosystem based on AES and RSA algorithms has been proposed.

Experiment results indicate that the pixel value distribution in the encrypted Meteosat images is even and uniform. The results have been analyzed thoroughly to study the strength of the confusion and diffusion properties, security and resistance level against some known attacks. Compared with other similar encryption schemes [18-20], our algorithm described above has higher security and can resist all kinds of known attacks

The proposed system is not just limited to this area, but can also be widely applied in the secure storage and transmission of confidential MSG images over the Internet and/or any shared network environment.

The tests have done in this study, and the obtained results are encouraged to focus the future research on new methods of integrity in the following areas of security to control integrity:

The marking (watermarking) as regards the insertion of a mark (watermark).

The IDC-hiding (hiding data) which is marked with a large amount of data.

The fingerprint is a form of marking where each object receives a known and unique identification number.

The digital signature is also a brand that simultaneously depends on the information obtained from the clear document and from hash functions.

REFERENCES

[1] Su-Yin Tan, “Meteorological Satellite Systems, Springer Briefs in Space Development,” 125 pp., 2014.

[2] W. Stallings, ‘’Cryptography and Network Security: Principles and Practice’’, 4th ed., Prentice Hall, 2011.

[3] Jonathan Katz, Yehuda Lindell, “Introduction to Modern Cryptography: Principles and Protocols”, ISBN: 978-1-58488-551-1, 2008.

[4] J. Daemen and V. Rijmen, “AES Proposal: The Rijndael Block Cipher,” tech. rep., Proton World Int.l, KatholiekeUniversiteit Leuven, ESAT-COSIC, Belgium, 2002.

[5] J. Daemen, V. Rijmen, “ the Design of Rijndael,” Springer Verlag, New York, Inc. Secaucus, NJ, USA, 2002.

[6] Wen, “AES encryption algorithm analysis and security stupy,” Computer Applications of Petroleum, Vol. 16 No.2, 2008.

[7] A. A. Shtewi, B. E. M. Hasan, A. El Fatah and A. Hegazy, "An Efficient Modified Advanced Encryption Standard (MAES) Adapted for Image Cryptosystems," IJCSNS International Journal of Computer Science and Network Security, Vol.10 No.2,pp.226-232 February 2010.

[8] B. Manoj andN Haribar Manula "Image Encryption and Decryption usingAES" International Journal of Engineering and Advanced Technology(IJEAT) ISSN: 2249 – 8958, Vol. 1, Issue5, June 2012.

[9] H. hamiche, M. lahdir , M. tahanout and S. djennoune , "masking digital image using a novel technique based on a transmission chaotic system and spiht coding algorithm" international journal of advanced computer science and applications (IJACSA), 3(12), 2012.

[10] W. Di-e, M. E. Hellman, "New Directions in Cryptography. IEEE Transactions on Information Theory," IT-22(6):644-654, 1976.

[11] R. L. Rivest, A. Shamir, and L. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Communications of the ACM, 21:120-126, 1978.

[12] S. Ammarah, V. Kaul, "Security Enhancement Algorithm for Data Transmission using Elliptic Curve Diffie - Hellman Key Exchange," International Conference & workshop on Advanced Computing 2014 (ICWAC 2014) – www.ijais.org.

[13] M. Dworkin, "Recommendation for Block Cipher Modes of Operation," NIST Special Publication 800-38A, 2001 Edition.

[14] W. Stallings, "cryptography and network security principles and practice," Prentice Hall Press Upper Saddle River, NJ, USA, 744 pp., 2011.

[15] R.Chakraborty, S. Agarwal, "Triple SV: A Bit Level Symmetric Block Cipher Having High Avalanche Effect,"nternational Journal of Advanced Computer Science and Applications, Vol. 2, No. 7, 2011.

[16] B. Parsharamulun, R. V. Krishnaiah, "A New Design of Algorithm for Enhancing Security in Bluetooth Communication with Triple DES," International Journal of Science and Research (IJSR), Volume 2 Issue 9, September 2013.

[17] S. Singh, S.K. Maakar, "A Performance Analysis of DES and RSA Cryptography," nternational Journal of Emerging Trends & Technology in Computer Science, Volume 2, Issue 3 May – June 2013.

[18] H. T. Panduranga and S. K. Naveen Kumar. "Hybrid approach for image encryption using SCAN patterns and Carrier Images." International Journal on Computer Science and Engineering 2.02 (2010): 297-300.

[19] J. Gao, ''.New Chaotic Image Encryption Algorithm Based on Hybrid Feedback''. Computer Application, 2008,28(2):434..436.

[20] G. Singh &S. Kinger,"Integrating AES, DES, and 3-DES Encryption Algorithms for Enhanced Data Security." International Journal of Scientific & Engineering Research 4.7 (2013): 2058.

AUTHOR PROFILE

Mohammed Belkaid Boukhatem was born in Algeria in 1987. He received his engineering degree in telecommunication from the the national institute of telecommunication and new technologies of information and communication Oran in 2010 and Magister degree in Electronics Remote Sensing from Mouloud Mammeri University of Tizi-Ouzou (Algeria) in 2015. He is currently pursuing his PhD thesis in LAMPA laboratory at Faculty of Electrical Engineering and Computing. His current research interests include image processing, signal processing and Meteosat image coding. His main application domain is cryptography.

Mourad Lahdir was born in Algeria in 1969. He received his Magister degree in Electronic from the Mouloud MAMMERI University of Tizi-Ouzou (Algeria) in 1999 and Ph.D. degree in Electronics Remote Sensing from the Mouloud MAMMERI University of Tizi-Ouzou in 2007. His research activities are image processing, Meteosat and hyperspectral image compression, wavelet and fractal image application, progressive data transmission and watermarking.

Mehdi Cherifi was born in Algeria in 1985. He received his engineering degree in Electronic from the Mouloud Mammeri University of Tizi-Ouzou in 2010 and Magister degree in Electronics Remote Sensing from Mouloud Mammeri University of Tizi-Ouzou (Algeria) in 2015.He is currently pursuing his PhD thesis in LAMPA laboratory at Faculty of Electrical Engineering and Computing. His current research interests include image processing, signal processing, Meteosat image compression.

On a Flow-Based Paradigm in Modeling and Programming

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

Abstract—**In computer science, the concept of flow is reflected in many terms such as data flow, control flow, message flow, information flow, and so forth. Many fields of study utilize the notion, including programming, communication (e.g., Shannon-Weaver communication model), software modeling, artificial intelligence, and knowledge representation. This paper focuses on two approaches that explicitly assert a flow-based paradigm: flow-based programming (FBP) and flowthing modeling (FM). The first is utilized in programming and the latter in modeling (e.g., software development). Each produces a diagrammatic representation, and these are compared. The purpose is to promote progress in a flow-based paradigm and its utilization in the area of computer science. The resultant analysis highlights the fact that FBP and FM can benefit from each other's methodology.**

Keywords—flow-based programming; conceptual description; data flow; flowthing model

I. Introduction

The notion of flow is quite ancient. The Greek philosopher Heraclitus (540–480 BCE) is known for a philosophy of flow, including his insight that one could not step twice into the same river, and "everything is always flowing in some respects" [1]. In this context, flow signifies a change with movement, and direction. It was viewed, metaphysically, as a universal principle, change, and the fundamental characteristic of nature. The notion of flow also appeared in China with Confucius (551–479 BCE), a contemporary of Heraclitus, whom he attributed with declaring that "everything flows like this, without ceasing, day and night" [2]. Accordingly, flow in some philosophical circles implies movement, change, and process [3] (see process philosophy, [4]).

This ancient concept of flow has been greatly discussed dialectically in many circles of philosophy, literature, and science (time flow, energy flow, information flow). Currently, it is a widely used concept in many fields of study. In economics, the goods circular flow model is well known; in management science, there is the supply chain flow. In computer science, the classical model of flow is the 1949 Shannon-Weaver communication model, representing electrical signal transfer from sender to receiver.

In computer programming, Flow-Based Programming (FBP) is a programming paradigm that uses a "data factory" metaphor for designing applications [5]. Other paradigms include Imperative, Functional, and Object-Oriented programming.

FBP utilizes networks of *black box* processes, which exchange data across predefined connections by message passing, where the connections are specified externally to the processes [5].

FBP is … a brand new way of thinking about application development, freeing the programmer from von Neumann thinking, one of the major barriers to moving to the new multiprocessor world, and has been evolving steadily over the intervening years. [6]

Recently, a new flow model (FM) has been proposed and used in several applications, including communication and engineering requirement analysis [7-11]. In FM, the flow of "things" indicates movement inside and between non-black box processes.

This paper focuses on these two approaches that explicitly assert that they adopt a flow-based paradigm: flow-based programming (FBP) and flowthing modeling (FM). The first is utilized in programming and the latter in modeling (e.g., software development). They are contrasted in terms of the diagrammatic representation each produces. The paper examines FBP and FM to find common concepts and differences between the two methodologies. Several advantages can be achieved from such a study:

- Enhancing of common concepts
- Identifying a foundation for tools and areas of application
- Furthering the development in use of the notion of flow

This would promote progress in a flow-based paradigm and its utilization in the area of computer science. After a review of background materials in section 2, section 3 explores some of the notions of FBP: Selector, Assign, Sequencizer, and Interactive network, in terms of FM representation. Section 4 discusses a specific problem: the "telegram problem" that is specified in FBP and then analyzed in FM.

II. Background Materials

As background information, subsection II.A differentiates between the two traditional mechanisms, data flow and control flow, with emphasis on data flow as the base of flow-based approaches. Subsections II.B and II.C summarize main ideas in FBP and FM. FM is covered more extensively because it is a less known approach. The FM example at the end of section II.C is a new contribution.

A. Data and Control Flow

In modeling and programming of software systems, structuring the relationships among processes (activities) described by two traditional mechanisms:

- Data flow, and
- Control flow, e.g., an execution order.

A data flow emphasizes data availability even within each task. In the FM version of this flow, data has the characteristic of *liquidity* (the state of being liquid). For example, according to Langlois [12], "Information is some sort of undifferentiated fluid that will course through the computers and telecommunications devices of the coming age much as oil now flows through a network of pipes." FM generalizes such a conceptualization to "anything that flows," i.e., is created, released, transferred, received, and processed.

Control flow gives the *execution order* of tasks in the form of instructions, e.g., sequences, branches, loops, and so forth. Conceptually, it is hard to think of a "control" that flows; rather, a more accurate description is to say that the instructions flow into the control sphere to be executed one after another, equivalent to typical sequential computing in the von Neumann model.

B. Flow-Based Programming

One of the important characteristics of FBP is the utilization of black box reusable modules, "much like the chips which are used to build logic in hardware" [13]. These black boxes, called components (see Fig. 1) are the basic building blocks used in constructing an application. "FBP is a graphical style of programming, and its usability is much enhanced by (although it does not require) a good picture-drawing tool" [13].

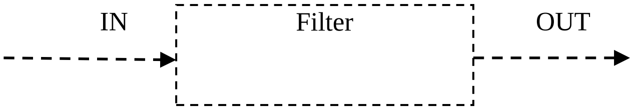


Fig. 1. Sample component in FBP (from [13])

The conventional approaches to programming (control flow) start with process and view data as secondary; business applications usually start with data and view the (data flow) process as secondary [13]. "Data" in FBP are atomic things and called "information packets" (or IPs). An Application is built up of many programs passing IPs around between them.

This is very like a factory with many machines all running at the same time, connected by conveyor belts. Things being worked on (cars, ingots, radios, bottles) travel over the conveyor belts from one machine to another on conveyor belts... [In a soft-drink bottling plant, you find] machines for filling the bottles, machines for putting caps on them and machines for sticking on labels, but it is the *connectivity* and the *flow* between these various machines that ensures that what you buy in the store is filled with the right stuff and hasn't all leaked out before you purchase it! [13] (Italics added)

FBP service requests have to do with communication between processes that include connections described in terms of *receive*, *send*, *drop*, *end of data*, … "IN" and "OUT" are called "ports" for receiving and sending IPs. They are doors that have an "inside" aspect and an "outside" aspect. A port is "a special place on the boundary through which input and output flow" [14]. A port establishes a relationship between the receives and sends inside the program, resembling subroutine parameters of function [13].

C. Flowthing Model

The Flowthing Model (FM) is also based on the notion of *flow*. It is a more model-oriented methodology.

Anybody having encountered the construction process will know that there is a plethora of flows feeding the process. Some flows are easily identified, such as materials flow, whilst others are less obvious, such as tool availability. Some are material while others are non-material, such as flows of information, directives, approvals and the weather. But, all are mandatory for the identification and modelling of a sound process. [15].

The word *flow* is rooted in the meaning "to move in a (steady) stream." The cognitive image of a liquid is therefore fused into every metaphor involving flow [16].

FM is used to develop a map of conceptual movement (analogous to the movement of blood through the heart and blood vessels) and states of *things* that are called *flowthings*. Goods, people, ideas, data, information, and money moving among *spheres* (e.g., places, organizations, machines …) are flowthings. Hence, the focus is not on information and information packets as it is in the case of FBP.

Flowthings flow in a non-black box system, called a *flowsystem*. The flowsystem is the "bed of a river" and the flowthing is the "water" that flows. It is a generalization of the input-process-output (IPO) model that has been used in FBP. A system is typically conceptualized as a set of interrelated constituents that collect (input), manipulate (process), and disseminate (output) data. The sequence of input-process-output is probably the most used *pattern* in computer science.

The basic IPO conception, used in FBP (exemplified in Fig. 1), is captured by "a process P acting on an input I and producing an output O" [17]. It views a system as a *black box* process with an interface, and the environment denotes everything outside that system. The interface can be invoked either by the system (output) or by the environment (input). The IPO notion of "process" hides structural divisions.

The FM *flowsystem* opens the black box by decomposing it into several specific (atomic/mutually exclusive) compartments and specifying flows within a system or a subsystem. *Flow* refers to the exclusive transformation of a flowthing passing among six states (also called stages) in a flowsystem: transfer (input/output), process, creation, release, arrival, and acceptance, as shown in Fig. 2. We use *receive* as a combined stage of *arrive* and *accept* whenever arriving flowthings are always accepted.

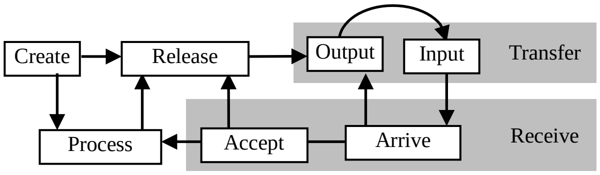


Fig. 2. Flowsystem

Each *stage* has its vocabulary:

- ***Create***: generate, appear (in the scene), produce, make, … In contrast to previous approaches, in FM *Creation* is considered a type of flow, i.e., from the sphere of nonexistence to the current sphere.
- ***Transfer***: transport, communicate, send, transmit … in which the flowthing is transported somewhere within or outside the flowsystem (e.g., packets reaching ports in a router, but still not in the arrival buffer).
- ***Process***: stage in which the form but not the identity of a flowthing is transformed, indicated by a seemingly endless choice of English verbs (e.g., compressed, colored, edited, marked, evaluated, ordered, …)
- **Released**: a flowthing is marked as ready to be transferred (e.g., airline passengers waiting to board)
- **Arrive**: a flowthing reaches a new flowsystem
- **Accepted**: a flowthing is permitted to enter the system

These stages are mutually exclusive; i.e., a flowthing in the *Process* stage cannot be in the *Created* stage or the *Released* stage at the same time. An additional stage of Storage can also be added to any FM model to represent the storage of flowthings; however, storage is a generic stage, because there can be *stored* processed flowthings, *stored* created flowthings, and so on.

The flowthings flow in specific "flow channels," changing in form and interacting with outside *spheres* (flowsystems in other systems), where *solid arrows* represent flows and *dashed arrows* represent triggering, e.g., receiving an action (e.g., a hit in the face) that triggers emotion (e.g., anger) that in turn triggers a physical reaction. Triggering may signify several semantics, including representing a flow. For example, in a case where a flowsystem triggers another flowsystem, it can indicate a signal flow, i.e., create a signal and send it to a destination flowsystem. When a *sphere* includes a single *flowsystem*, then only one box is drawn to represent both the sphere and its flowsystem.

Example: In mathematics, a function *f*(x) takes an input x and returns an output f(x). In teaching the concept of function, one metaphor describes function as a "black box" that for each input returns a corresponding output [18] (see Fig. 3). A function is described as the set of rules that convert the input to output, analogous to the work of a machine.

This approach can be applied to illustrate the Big-O Notation used in elementary computer science courses. It has been found that students have difficulty understanding the definition and the method of finding the Big-O for a given function.

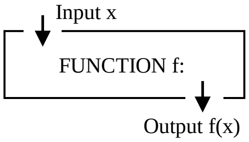


Fig. 3. Black box representation of functions

For example, a textbook [19] used in that context defines the Big-O as follows:

Let f and g be functions from the set of integers or the set of real numbers to the set of real numbers. We say that f(x) is O(g(x)) if there are constants C and k such that whenever x > k.

Using FM (Fig. 4), the two functions f(x) and g(x) (circles 1 and 2, respectively, in Fig. 4) can be viewed as *spheres* with x as a *flowthing* (circle 3 in Fig 4 – (a)).

The FM description highlights searching for **k**, **g(x)**, and **C** as follows:

- Select k such that *x > k* (circle 4),
- Select g(x)
- Select constant C (5)
- Multiply g(x) by C such that we produce Cg(x) > f(x) (6).

Values of *x > k* flow to *f(x)* and the selected *g(x)*. Both *f(x)* and *Cg(x)* flow to create ***f(x) = O(g(x))*** if ***Cg(x) > f(x)***. Fig 4 (b) illustrates finding the Big O for *x2 + 2x + 1*. In this case, the students keep selecting k, g(x) (a *minimum* function is the best), and C, in the FM depiction, as an educational game. This visual representation helps in finding k = 1, g(x) = x^2 (minimum), and C = 4 to satisfy Cg(x) > f(x).

Note how the variables x, functions, and requirement (Cg(x) > f(x)) are represented uniformly, as flowthings and spheres.

III. Contrasting Diagrammatic Representations of FBP and FM

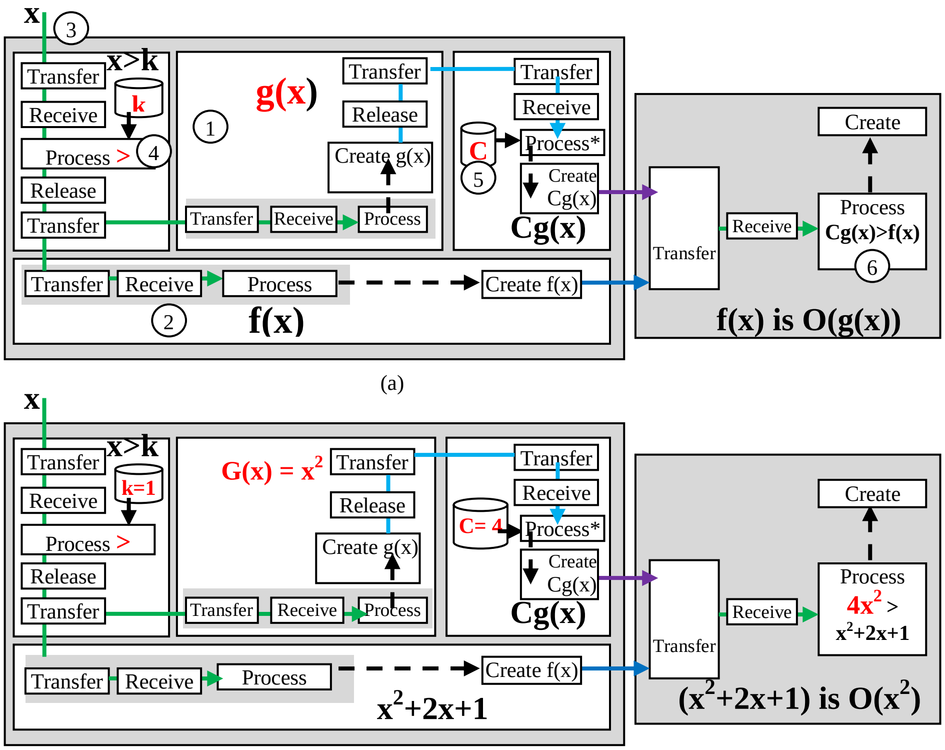
This section explores some of the features of FBP and FM as part of the attempt to bring their diagramming methodologies into closer alignment, possibly advancing the flow-based paradigm in its different forms for programming and modeling.

A. Selector

Fig. 5 shows a sample component call *Selector* in FBP. It applies some criterion "c" to all incoming IPs, and sends out the ones that match the specified criterion while sending the rejects to the other output port (REJ) [13].

Here, we can identify a basic difference between FBP and FM: conceptually, from the FM point of view, the output in this component is a different type of flowthing from the input. It is analogous to a currency handler who receives banknotes and then separates them by currency, say, by dollars and pounds. Accordingly, in the FM description of the selector (Fig. 6), each flow is represented as a separate stream.

One basic FM principle is that different types of element flows are not mixed in the same diagram, eliminating ambiguity and difficulty in identifying the streams of flow. This mixing of flows is a basic engineering assumption, for example, the semantics of the arrows where different flows are intermixed, analogous to representing electrical lines and water pipes by the same arrow in the blueprint of a building. Figs. 7 show the corresponding pseudocodes in FBP and FM.



(a)

(b)

Fig. 4. Using an FM diagram to illustrate and find the Big O

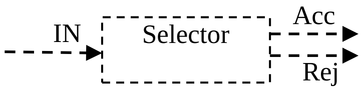


Fig. 5. Sample component in FBP (from [13])

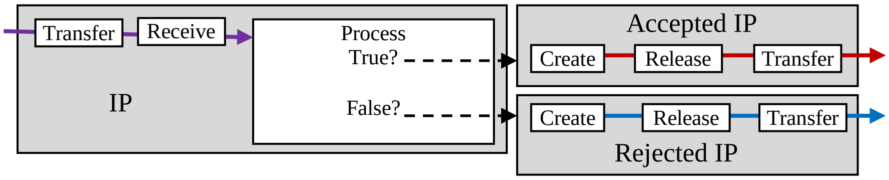


Fig. 6. IP is processed to generate one of two types of IPs

```
receive from IN using a
do while receive has not reached end of data
       if c is true
           send a to ACC
      else
            send a to REJ
       endif
       receive from IN using a
enddo
```

```
IP
  Transfer/receive
  Process: If True? then Accepted IP
             If False? then rejected IP
Accepted IP
   Create
   Release/Transfer
Rejected IP
   Create
   Release/Transfer
```

Fig. 7. Pseudocode in FBP (*left* [13]) and FM (*right*)

B. *Sequencizer*

Consider the component called "Sequencizer" used "in all existing FBP systems which simply accepts and outputs all the IPs from its first input port element, followed by all the IPs from its second input port element, and so on until all the input port elements have been exhausted" [13] (see Fig. 8).

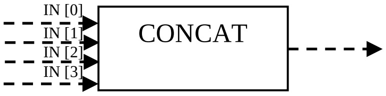


Fig. 8. Diagrammatic representation of CONCAT in FBP (from [13])

A Sequencizer is often used to force a sequence of data being randomly generated from a variety of sources, e.g., IPs generated by different processes that can then be printed out in a fixed order in a report. To simplify, we can understand the Sequencizer in terms of, say, numbers, e.g., "123" and "567", that are concatenated, as into a sequence, e.g., "123 567". In this case, conceptually, the sequence is a different flowthing from its constituents; thus it has its own flowsystem in CONCAT, as shown in Fig. 9. Accordingly, "opening" the black box, a notion that has been adopted by FBP, reveals not only different internal processes, but also the structure of the component.

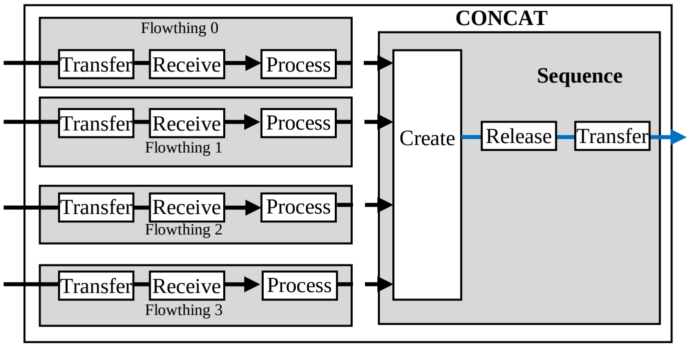


Fig. 9. CONCAT in FM

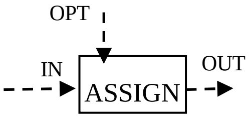


Fig. 10. Diagrammatic representation of ASSIGN in FBP

C. *Assign*

The *Assign* component in FBP "simply plugs a value into a specified position in each incoming IP, and outputs the modified IPs" [13]. OPT receives the specification of where in the incoming IPs the modification is to take place, and what value is to be put there (see Fig. 10). The FM representation of *Assign* is shown in Fig. 11. The arrows are drawn in different colors to emphasize different flows.

According to Morrison [13], to tell the black box *Select* component which fields to select, in FBP, the application designer specifies this information through a mechanism called an *Initial Information Packet* (IIP). For example, in the Selector component discussed previously, the selection criteria (true and false or any other values) can be fed to the *Selector* along with other criteria (similar to OPT in Fig. 10).

Fig. 12 shows the FM representation of this structure. IPs and OPT are input, and an IP/OPT flowsystem (circle 1) is a system that deals with a type that is a supertype of IP and OPT (2 and 3, respectively), analogous to fixing types in, say, C++. In Fig. 12, the process triggers the creation of accepted and rejected IPs (4 and 5, respectively).

D. *Interactive network*

An interactive network is a general schematic (see Fig. 13) in which requests coming from users enter the diagram, and responses are returned. The "back-ends" communicate with systems at other sites. The cross-connections are requests that do not need to go to the back ends, or that must cycle through the network more than once before being returned to the user [13].

Fig. 13 is conceptually disturbing because the cross-connections mix flows of requests and responses. Imagine mixing the ingoing/outgoing pipes in engineering projects.

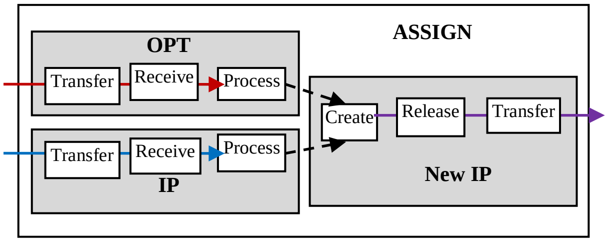


Fig. 11. Diagrammatic representation of ASSIGN in FM

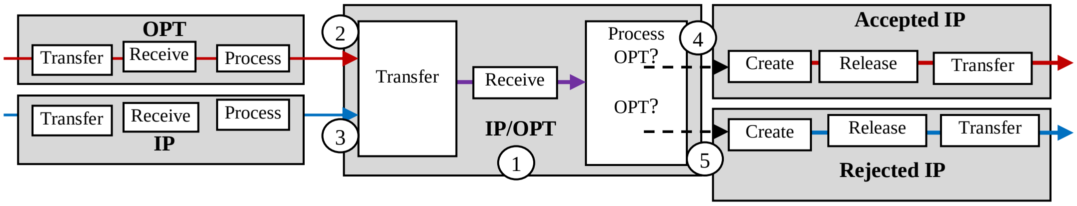


Fig. 12. Select component where the criteria of decision is input

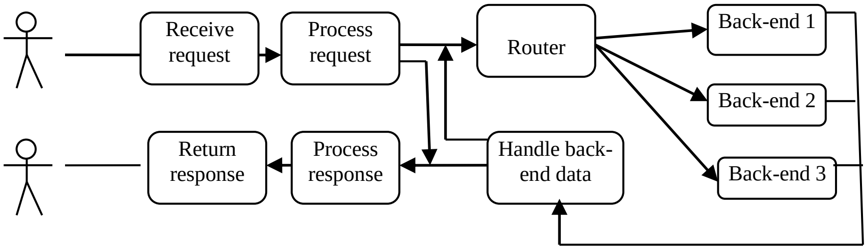


Fig. 13. Simple interactive network (redrawn from [13])

Accordingly, Fig. 14, which shows the FM representation, ignores these cross-connections. In the figure, the user creates a request (circle 1 in the figure) that flows to the system (2), where it is processed (3). Then it flows to the router (4) and is processed (5) and sent to one of the back-ends (6). In the back-end, the request is processed to trigger (7) the creation (8) of a response. The response flows to the Handle back-end data (9) where it is processed (10), then sent to the return module (11) that sends it to the user (12). For simplicity sake, cross-connections are ignored in Fig. 14. It is possible to handle them by capturing such requests in *process* when they flow in the *Receive Request* flowsystem and then treat them separately.

The flows of requests and responses are separated in the FM representation. It seems that a definition of flow is lacking from flow-based programming. In FM, a flow refers to the movement of flowthings among stages and spheres. A flowthing is a thing that can be created, released, transferred, received, and processed. It has its own stream of flow. If flow types are mixed, this is performed explicitly, in a flowsystem that represents their supertype, e.g., *integers* and *reals* are handled by the flowsystem *number*.

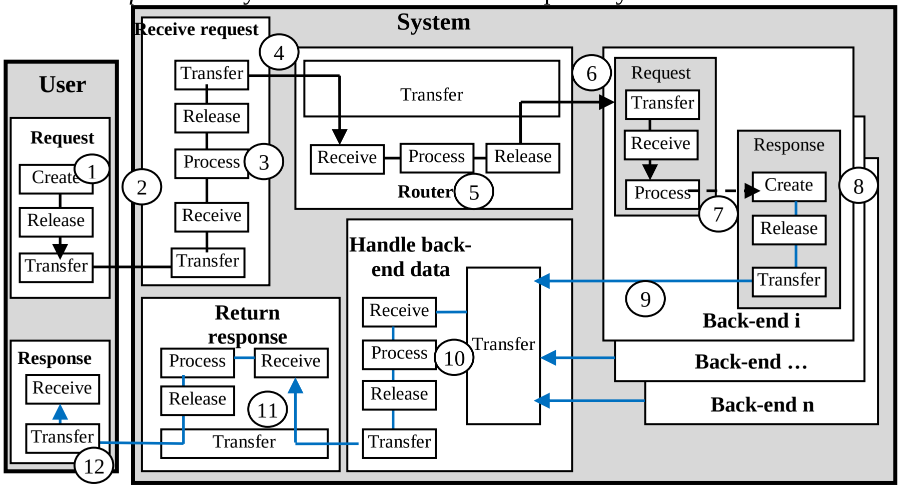


Fig. 14. Simple interactive network in FM

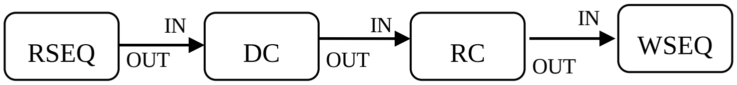


Fig. 15. FBP diagrammatic representation of the telegram problem (redrawn, partial from [13])

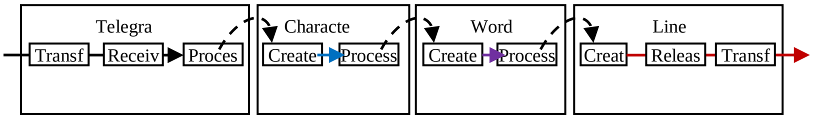


Fig. 16. FM diagrammatic representation of the telegram problem

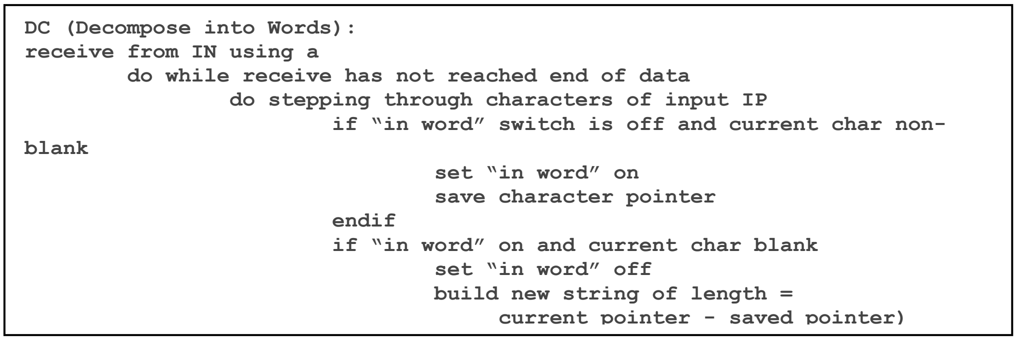
IV. Programming Aspects

This section discusses a specific problem: the “telegram problem” that is specified in diagrammatic representation and textual format. In FBP, the “black box” (component) seems to be specified anew, with respect to the explicit flow-based high-level diagram. In FM, the details of flow and its stages involve just a refinement to the FM depiction.

Consider a simplified telegram problem [20] in which a program is required to process telegrams. Each telegram is available as a sequence of words and spaces. Telegrams are to be processed to become output with all but one space between words eliminated. In FBP, words are treated as IPs. Fig. 15 shows the FBP description of the solution where RSEQ means “Read Sequential”, WSEQ means “Write Sequential”, DC is “DeCompose”, and RC is “ReCompose”. Fig. 16 shows the corresponding FM depiction.

Contrasting the two representations, we see the difference in terms of retracing of components by a flowsystem. The flowsystem expresses the *type* of flowthing and the basic *operations* performed on it, e.g., create, process, …, in addition to defining a flow in terms of flowthings. FBP represents all types of flow with a solid arrow, implicitly relating the type of flow to the component that outputs it.

As mentioned previously, *triggering* in FM may have several semantics, including representing a flow. In Fig. 16, triggering causes *creation* in the next flowsystem. For example, Telegram (sphere) represents the flowsystem of a *string* (flowthing) (remember that when a sphere includes a single flowsystem, both are represented by one box). The triggering causes the appearances (creation) of *characters* in the Character sphere. Clearly, Telegram “slices” the “string” into “characters” and sends them to Character. So, why do we put **Create** in the Character sphere? From a purely semantical point of view, Telegram does not know *character*. It is – from the point of view of Telegram – a collection of processed pieces of a string. This is similar to representing string as an array of characters in C++. These “pieces of string” are then shipped to Character. So, *triggering* (in this case) means the flow of these pieces from String (where they are pieces of string) to *Character* (where they are recognized as characters).



```
DC (Decompose into Words):
receive from IN using a
        do while receive has not reached end of data
                do stepping through characters of input IP
                        if "in word" switch is off and current char non-
blank
                                set "in word" on
                                save character pointer
                        endif
                        if "in word" on and current char blank
                                set "in word" off
                                build new string of length =
                                     current pointer - saved pointer)
```

Fig. 17. FBP programming of the telegram problem (partial from [13])

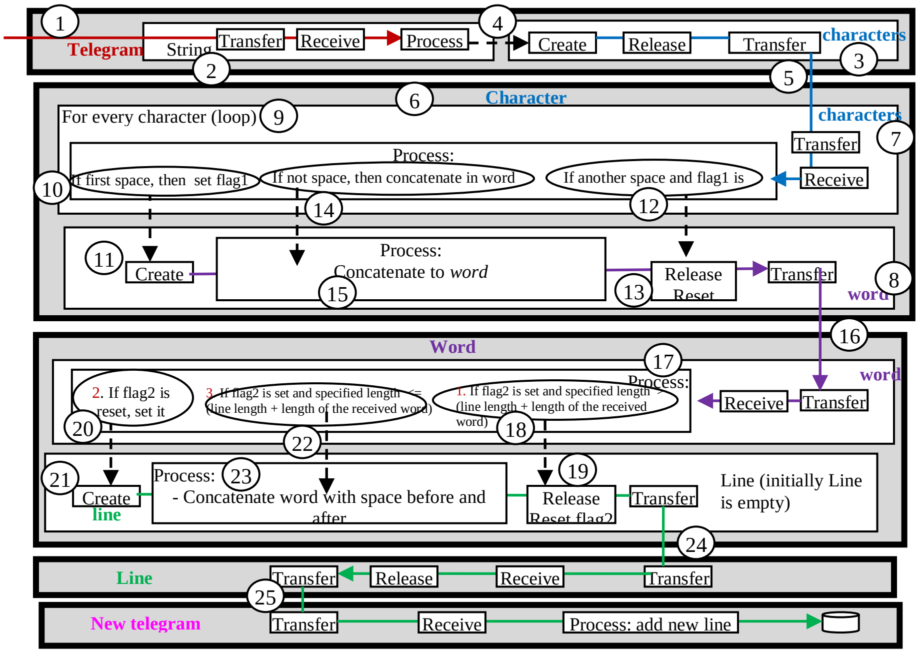

1
Telegram
String
2
Transfer
Receive
Process
4
Create
Release
Transfer
characters
3
5
6
Character
For every character (loop)
9
characters
7
Transfer
Process:
10
If first space, then set flag1
If not space, then concatenate in word
If another space and flag1 is
Receive
14
12
11
Create
Process:
Concatenate to word
15
13
Release
Reset
Transfer
8
word
16
Word
17
Process:
word
2. If flag2 is reset, set it
3. If flag2 is set and specified length <= (line length + length of the received word)
1. If flag2 is set and specified length > (line length + length of the received word)
Receive
Transfer
20
22
18
21
Create
line
Process:
23
- Concatenate word with space before and after
19
Release
Reset flag2
Transfer
Line (initially Line is empty)
24
Line
Transfer
Release
Receive
Transfer
25
New telegram
Transfer
Receive
Process: add new line

Fig. 18. FM programming diagram of the telegram problem

However, for simplicity's sake, in a more elaborate specification, we will assume that in triggering of this type, the *creation* happens in the *source* flowsystem, after which the flowthings flow to the destination flowsystem of triggering. For example, in Fig. 16, *words* are created (from characters) in Character and flow to the Word sphere.

So, in moving to programming, Fig. 17 shows the FBP pseudocode for component DC (see Fig. 15).

It seems that the "black box" DC has an extensive interior. What is disturbing in the FBP implementation is the lack of explicit connection between the diagrammatic and the pseudocode representations. Where is the flow in the pseudocode? Is it, implicitly, in the design of the so-called Ports, IN and OUT?

In FM, the effort to realize a diagrammatic representation is easily facilitated, as shown in Fig. 18, in a simplified telegram problem where we assume that there are no multiple spaces. The telegram flows (circle 1) in its sphere, which includes two flowsystems, *string* (2) and *character* (3), drawn according to our previous explanation of the semantics of *triggering* Create (creating characters in the telegram sphere). The processing of the telegram as a string triggers (4) the creation of characters (e.g., string in C++ becomes an array of characters). The characters flow to the *Character sphere* (6), which includes *characters* and *word* flowsystems (7 and 8, respectively). There is also a declaration of a loop for all incoming characters (9). A loop is also a type of sphere in FM. Depending on character processed, a decision is made:

- If it is a first space, then set flag1 and create an empty word (10 and 11).
- If another space and flag1 is set (initially reset), then release the word (12) and reset tflag1 (13). Note that the created *word* flows to be processed and released.
- If not space, then concatenate it to the end of currently built *word* (14 and 15, respectively)

Again, this mixing of character and word flowthings is done to simplify the diagram, as discussed previously. For Character, it is a matter of "padding up" characters, not creating *words*.

The *words* flow (16) to the Word sphere, where they are processed (17) to construct a *line*.

- (a) If flag2 is set and specified length > (line length + length of the received word), then release *line* and rest flag2 (18, 19)
- (b) If flag2 is reset, set it and a create new *line* – initially flag2 is reset (20, 21)
- (c) If flag2 is set and specified length <= (line length + length of the received word) then pad *word* to *line* (22, 23)

These *if statements* may need some synchronization, e.g., a new word waits to output previous *line* and reset flag2. Accordingly, the *line* flows to the Line sphere (because it includes a single flowsystem, it is drawn as such - 24), then to the New Telegram sphere (25).

Fig. 19 shows the textual pseudocode after removing Transfer and Release stages.

The point in this type of description is to demonstrate FM systematic refinement along a *flow-base* conceptualization. Levels of detail follow the same rhythm of flow, in contrast to an eruption that opens the "black box" in such a way that the flow mostly vanishes.

V. Conclusion

This paper has focused on two approaches that explicitly assert that they adopt a flow-based paradigm: flow-based programming (FBP) and flowthing modeling (FM). Extensive literature has been published on FBP dating back to the nineties of the last century. FM is much more recent and has not been utilized in programming. The resultant analysis indicates that FBP and FM can benefit from each other's methodology. It seems that FBP can benefit from the theoretical ideas in FM, while FM can be improved by considering the rich programming efforts in FBP. Both can promote development in a flow-based paradigm and its utilization in computer science.

Future work will explore the possibility of enhancement of the programming aspects in FM utilizing proven notions of FBP [21].

```
(All flags are initially reset, only one space between words)
Telegram
Process:
        Trigger Create character
Characters
For every character (loop)
Process:
          If first space, then set flag1, Create word
          If not space, then concatenate in word
          If another space and flag1 is set, trigger releasing word, reset flag1
Words
Process:
             1.     If flag2 is set and specified length  > (line length + length of the received word) trigger release
                    line, and reset flag2
             2.      If flag2 is reset, set it, and trigger Create new line
             3.     If flag2 is set and specified length  <= (line length + length of the received word) trigger
                    padding word to line
Line (assumed that length is given)
Release line to New telegram
New telegram (assumed initially empty)
      Process (add to the new telegram)
```

Fig. 19. FM pseudocode of the telegram problem

REFERENCES

[1] W. Daniel, and D. W. Graham, Heraclitus. In: Stanford Encyclopedia of Philosophy (2011), . http://plato.stanford.edu/entries/heraclitus/

[2] A. N. Beris, and A. J. Giacomin, πάντα ῥεῖ (Everything Flows): Motto for Rheology, Polymers Research Group Technical Report Series QU-CHEE-PRG-TR--2014-3 (May 23, 2014), http://hdl.handle.net/1974/12193.

[3] D. Chen, Metaphorical Metaphysics in Chinese Philosophy: Illustrated with Feng Youlan's New Metaphysics. Lexington Books, Lanham, MD (2011).

[4] Process Philosophy. In: Stanford Encyclopedia of Philosophy (Oct 15, 2012), http://plato.stanford.edu/entries/process-philosophy/

[5] J. P. Morrison, Flow-based Programming, http://www.jpaulmorrison.com/fbp/

[6] Flow-Based Programming, theTrendyThings blog (January 10, 2015), http://thetrendythings.com/read/17922

[7] S. Al-Fedaghi, States and Conceptual Modeling of Software Systems. Int. Rev. Comput. Softw. 4(6), 718–727 (2009).

[8] S. Al-Fedaghi, Developing Web Applications. Int. J. Softw. Eng. Appl. 5(2), 57–68 (2011).

[9] S. Al-Fedaghi, Conceptualization of Various and Conflicting Notions of Information. Inform. Sci. 17, 295–308 (2014).

[10] S. Al-Fedaghi, An Alternative Approach to Multiple Models: Application to Control of a Production Cell. Int. J. Control Automat. SCOPUS 7(4) (2014).

[11] S. Al-Fedaghi, Information System Requirements: A Flow-Based Diagram versus Supplementation of Use Case Narratives with Activity Diagrams. Int. J. Bus. Inform. Syst. 17(3), 306–322 (2014).

[12] R. Langlois, Systems Theory, Knowledge and the Social Sciences. In Machlup, F., Mansfield, U. (eds.) The Study of Information: Interdisciplinary Messages, pp. 581-600. Wiley, New York (1983).

[13] J. P. Morrison, Flow-Based Programming: A New Approach to Application Development. Van Nostrand Reinhold, New York, (1994). ISBN 0-442-01771-5. http://cs-wwwarchiv.cs.unibas.ch/lehre/fs08/cs506/_Downloads/book.pdf

[14] G. M. Weinberg, An Introduction to General Systems Thinking. John Wiley and Sons, New York (1975).

[15] R. Stevens, P. Brook, P., K. Jackson, and S. Arnold, Systems Engineering: Coping with Complexity. Prentice Hall PTR (1998).

[16] J. D. Casni, 'Flow' Hits Its Peak. Blog entry, http://metaphorobservatory.blogspot.com/2005/11/flow-hits-its-peak.html

[17] D. Gile, Opening Up in Interpretation Studies. In: Snell-Hornby, M., Pöchhacker, F., Kaindl, K. (eds.) Translation Studies: An Interdiscipline, pp. 149–158. John Benjamins, Amsterdam (1994).

[18] Boundless.com. Functions and Their Notation. In: Boundless Algebra. Jan. 25, 2015. Retrieved Apr. 13, 2015 from https://www.boundless.com/algebra/textbooks/boundless-algebra-textbook/graphs-functions-and-models-2/functions-an-introduction-17/functions-and-their-notation-98-5828/

[19] K. H. Rosen, Discrete Mathematics and Its Applications, 7th ed. (2011). ISBN: 0073383090.

[20] J. M. Barzdin, A. A. Kalnins M. I. Auguston, SDL Tools for Rapid Prototyping and Testing. In: Faergemand, O., Marques, M.M. (eds.) SDL'89: The Language at Work, pp. 127–133. North-Holland (1989).

[21] J. P. Morrison, Flow-Based Programming, 2nd Edition: A New Approach to Application Development (2010). ISBN-10 1451542321.

(AMDC) Algorithm for wireless sensor networks in the marine environment

Rabab J. Mohsin
School of Computer Science and Electronic Engineering
University of Essex
Colchester , Essex CO4 3SQ, UK
AL-Mustansiriya University,College of Engineering
Computer and Software Engineering Dept.,Baghdad, Iraq

John Woods , Mohammed Q. Shawkat
School of Computer Science and Electronic Engineering
University of Essex
Colchester , Essex CO4 3SQ, UK

***Abstract*—Data compression is known today as one of the most important enabling technologies that form the foundation of the majority of data applications and networks as we know them, including wireless sensor networks and the popular world wide net (internet). Marine data networks are gaining increasing interest in the research community due to the increasing request for data services over the sea. There are a very narrow range of available solutions because of the absence of infrastructure over such vast water surfaces. We have previously proposed applying MANET networks in the marine environment using VHF technology available on the majority of ships and vessels in order to gather different sensor data such as sea depth, temperature, wind speed and direction, etc. and send it to a central server to produce a public information map. We also discusses the gains and drawbacks of our proposal including the problem of low rate data transmission offered by VHF radio limited to 9.6 Kbps. In this paper we investigate the application of appropriate data quantization and compression techniques to the marine sensor data collected in order to reduce the burden on the channel links and achieve better transmission efficiency.**

Keywords—Wireless sensor network, Mobile Ad hoc Network, Very High Frequency, Sensor.

I. Introduction

In wireless senor networks, deployed sensor nodes periodically collect readings and send them to sinks (or base stations) via wireless channels, (WSNs) are resource constrained : limited power supply, low bandwidth for communication, processing speed, and memory storage. WSNs are suitable for large scale data collection purposes in different situations such as environmental monitoring, habitat monitoring, surveillance, structural monitoring, equipment diagnostics, disaster management, and emergency response [1]

Sensor nodes in WSNs are usually self-organized and they communicate with each other in a wireless manner to perform a common task. The nodes are generally deployed in large numbers and distributed randomly in an ad-hoc manner in the sensor field. Each node is equipped with battery, wireless transceiver, microprocessors, sensors, and memory. upon deployment, the sensor nodes form a network through short-range wireless communication. The collected data by each sensor node is transmitted wirelessly to the sink either directly or through multihop communication [2].

Oceans have abundant resources, wide spaces and play important roles in the activities of the Earth's environment and climate [3]. Oceanography is very rich, involving marine physics, marine chemistry, marine biology, marine geology and many other research fields [4]. How to collect data effectively to understand the marine environment, so as to exploit marine resources, has become one of the most important technologies in the oceanic areas [3].

Marine sensor data come from sensor networks deployed in a marine environment. Types of marine environments include rivers, seas and oceans. In most cases, the raw data stored in databases are first retrieved and processed using mathematical and statistical tools and are then visualized dependent on the user requirements.

In this paper, we categorize the more important sensor data to be gathered by ships, we analyse the datas characteristics in terms of sensor reading range and acceptable decimal place accuracy. We then employ this study to obtain a quantization and compression algorithm by using our model (Average Marine Data Compression (AMDC)) to reduce the traffic size on our low data rate VHF channel proposed in [3] for MANETs in the marine environment. We evaluate and compare the proposed data compression techniques with other known techniques and evaluate its suitability for deployment on resource constrained devices such as a WSN node in the marine environment.

Effective marine data processing and transmission is very important for facilitating marine environmental studies. Several works have been conducted in this area and below we summarize the more relevant ones. In [4] the authors present a data prediction model calculated from the latest three values acquired. From these values, the proposed algorithm calculates the lowest, the highest, and the medium value. At the end it transmits the difference between one of the calculated values and the actual one, depending on its position from these. In [5] the authors have described a variation of the lossless LZW algorithm relating to the common sensor platforms with a few kilobytes of memory. This version can achieve the compression of a data block with a length of 528 bytes at a time.

In [6] the authors examine the utility of linear predictive coding in reducing the amount of data storage required for signals gathered in ocean bottom seismology. In this study, a set of 12 typical signals were repeatedly encoded with the storage allocated decreasing from an initial 12 bits per datum to 2. The error introduced was then compared to the performance achieved by simply rounding off the lowest bits of the data, to estimate the rate distortion limit. It was found that this scheme consistently introduced about 15 times (4 bits) less distortion both in terms of the root-mean-square (rms) error and in terms of the maximum error than rounding the data.

In [7] the authors discussed the spatial and statistical characteristics of underwater imagery that facilitate compression by well-known algorithms such as JPEG, vector quantization (VQ), and visual pattern image coding (VPIC). They considered statistical distributions of target and background grey levels obtained from truthed imagery, as well as power spectral analysis of target-background differences. The former measures facilitate parameter selection in VQ and VPIC, while the latter are important in JPEG.

In [8]the authors designed a wavelet based hybrid video encoder which employs entropy-constrained vector quantization (ECVQ) with overlapped block-based motion compensation. The ECVQ codebooks were designed from a statistical source model which describes the distribution of high sub band wavelet coefficients in both intra frame and prediction error images. Results indicate that good visual quality can be achieved for very low bit-rate coding of underwater video with the proposed algorithm.

II. Digital Signal Processing

The basic communication problem may be posed as conveying source data with the highest fidelity possible without exceeding an available bit rate, or it may be posed as conveying the source data using the lowest bit rate possible while maintaining a specified reproduction fidelity. In either case, a fundamental trade-off is made between bit rate and signal fidelity. The ability of a source coding system to suitably choose this trade-off is referred to as its coding efficiency or rate distortion performance. To represent a signal in the digital domain, it has to go through a number of steps as shown in Figure(1) which are described in turn [9].

A. Sampling [10]

A digital signal is formed from an analogue signal by the operation of sampling, quantizing, and encoding. The analogue signal, denoted x(t), is continuous in both time and amplitude. The result of the sampling operation is a signal that is still continuous in amplitude but discrete in time. Such signals are often referred to as sampled-data signals. A digital signal is formed from a sampled data-signal by encoding the time-sampled values onto a finite set of values.

B. Quantization

Quantization is the division of a quantity into a discrete number of small parts, often assumed to be integral multiples of a common quantity. The oldest example of quantization is rounding off, which was first analysed by Sheppard [11] for the application of estimating densities by histograms [12]. Quantization makes the range of a signal discrete, so that the quantized signal takes on only a discrete, usually finite, set of values. Unlike sampling (where we saw that under suitable conditions exact reconstruction is possible), quantization is generally irreversible and results in loss of information. It therefore introduces distortion into the quantized signal that cannot be eliminated. One of the basic choices in quantization is the number of discrete quantization levels to use. The fundamental trade-off in this choice is the resulting signal quality versus the amount of data needed to represent each sample [13].

C. Encoding

Encoding is a digital symbol processing operation in which the digital form of the information is changed for improved communication. In general, encoding contains many different processes, such as ciphering, compression, and error control coding. One of the main purposes of encoding is compressing information. By using data compression we can reduce the disk space needed to store data in a computer. In the same way we can decrease the required data rate on the line to a small fraction of the original information data rate. We could, for example, use very short codes for the most common characters instead of the full seven-bit ASCII code. Rarely needed characters would use long codes and the total data rate would be reduced [14].

III. data compression

Data compression is the art of reducing the number of bits needed to store or transmit data. There are two types of compression, lossy and lossless. Lossy compression reduces file size by eliminating some unneeded data that will not be recognized by the human after decoding, this is often used in video and audio compression. Losslessly compressed data on the other hand, can be decompressed to exactly its original value. This is important because if a file is lost even a single bit after decoding, will mean the file is corrupted [15].

These steps can be used to reduce the transmission overhead attributed to data transmission. WSN devices are universal and applicable to many sensing and control applications, making the characteristics of various presented datasets wide and varied.

IV. Arithmetic code

Arithmetic coding is a technique for coding that allows the information from the messages in a message sequence to be combined to form a single bit stream. A code word is not used to represent a symbol of the text. Instead it uses a fraction to represent the entire source message [15].The technique allows the total number of bits sent to asymptotically approach the sum of the self information of the individual messages (recall that the self information of a message is defined as $log_2(1/P_i)$).

In the following discussion we assume the decoder knows when a message sequence is complete either by knowing the length of the message sequence or by including a special end-of-file message. We will denote the probability distributions of a message set as p(1), . . , p(m), and we define the accumulated probability for the probability distribution as in equation 1

$$f(j) = \sum_{i=1}^{j-1} p(i)(j = 1, ..., m). \quad (1)$$

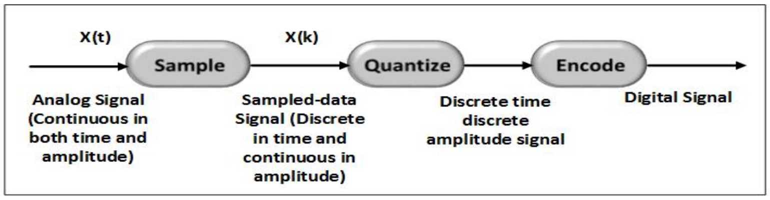

X(t)
Sample
X(k)
Quantize
Encode
Analog Signal (Continuous in both time and amplitude)
Sampled-data Signal (Discrete in time and continuous in amplitude)
Discrete time discrete amplitude signal
Digital Signal

Fig. 1: Sampling, quantizing, and encoding.

The main idea of arithmetic coding is to represent each possible sequence of n messages by a separate interval on the number line between 0 and 1.The occurrence probabilities and the cumulative probabilities of a set of symbols in the source message are taken into account. The cumulative probability range is used in both compression and decompression processes. In the encoding process, the cumulative probabilities are calculated and the range is created in the beginning. While reading the source character by character, the corresponding range of the character within the cumulative probability range is selected. Then the selected range is divided into sub parts according to the probabilities of the alphabet. Then the next character is read and the corresponding sub range is selected. In this way, characters are read repeatedly until the end of the message is encountered. Finally a number should be taken from the final sub range as the output of the encoding process. This will be a fraction in that sub range. Therefore, the entire source message can be represented using a single fraction. To decode the encoded message, the number of characters of the source message and the probability/frequency distribution are needed [16] .

V. Benefits of the Proposed System

Our proposed work is a real world environmental sensor application surveillance system in a marina environment. The purpose of our network is to collect environmental information from different ships. Each ship has a box for AIS and a VHF transceiver. A number of sensors will be placed on the ship to get useful information of(Position,Velocity,Humidity, Temperature, Wind speed, Wind direction, Barometric Pressure, Salinity,Depth, and PH). This data is then sent through a mobile ad-hoc network of ships by multi hop over VHF radio to a destination computer where the accumulated collected data, can be processed for end user applications (Accuracy of Weather information, up to date depth information, and etc.). Because of the low bandwidth available , it is a beneficial for WSNs to employ data compression algorithms. Low-complexity and small size data compression algorithms for sensor networks are therefore essential. The proposed algorithm; Average Marine Data Compression (AMDC) solves this problem by reducing the amount of data presents in the network channel.

VI. Summarization and Analysis of Marine Sensor Data

The most important sensors applied in our proposed sensor network are as follows:

A. Position

Any location on Earth is described by two numbersits latitude and its longitude. If a ship wants to specify position on a map, these are the coordinates they would use. Actually, these are two angles, measured in degrees, minutes of arc and seconds of arc. These are denoted by the symbols (, ,) e.g. 35 43 9 means an angle of 35 degrees, 43 minutes and 9 seconds. A degree contains 60 minutes of arc and a minute contains 60 seconds of arcand you may omit the words of arc where the context makes it absolutely clear that these are not units of time. Calculations often represent angles by small letters of the Greek alphabet, and that way latitude will be represented by l (lambda, Greek L), and longitude by f (phi, Greek F) [17].

B. Velocity

Velocity is the rate of change of the position of a ship, equivalent to a specification of its speed and direction of motion e.g. (60 km/h to the north). The applicable range in the marine environment would be between 0 and 75 Km/h.

C. Humidity

Humidity of air is a function of both water content and temperature. The relative humidity of an air-water mixture is defined as the ratio of the partial pressure of water vapour (H2O) in the mixture to the saturated vapour pressure of water at a given temperature. The applicable range in the marine environment would be between 0 and 100 %.

D. Temperature

Temperature is a comparative objective measure of hot and cold. It is measured, typically by a thermometer, through the bulk behaviour of a thermometric material, detection of heat radiation, or by particle velocity or kinetics. It may be calibrated in any of various temperature scales, Celsius, Fahrenheit, Kelvin, etc. The applicable range of air temperature in the marine environment would be between -50° and 50° C. While sea water temperature is inclusive between -2° and 36° C.

E. Wind Speed

Wind speed is the measure of motion of the air with respect to the surface of the earth covering a unit distance over time. The applicable range would be between 0 and 110 mph.

F. Wind direction

Wind direction is an indicator of the direction that the wind is heading and is usually measured in a degree between 0 and 360.

G. Barometric pressure

Barometric pressure (also known as atmospheric pressure) is the force exerted by the atmosphere at a given point. It is known as the weight of the air. A barometer measures barometric pressure. Measurement of barometric pressure can be expressed in millibars (mb) or in inches or millimetres of mercury (Hg). The applicable range would be between 800 and 1100 mb.

H. Salinity

Salinity precisely measures the total dissolved salt content of ocean or brackish water. The applicable range would be between 0 and 44%.

I. Depth

A depth sensor measures sea level close to the shore and in the deep ocean. The highest applicable reading in the marine environment is about 10,925 m.

J. PH

A PH sensor measures sea and ocean water acidity in the range between 0 and 14 . A neutral reading would be around 7 .

For each of the sensors mentioned previously we have set the extreme lower and upper limits of the sensors readings likely to be found in the marine environment as well as the level of accuracy required to represent each reading. This would enable us during the quantization process to reduce the number of bits required for representing the readings of each sensor limited to the predefined ranges and accuracy steps within those ranges. Table II below shows each sensor measurement and the corresponding step level required.

VII. Quantization of Marine Sensor Data

All the bit calculations were done according to the quantization rules in a straightforward manner where we use the range of readings for each sensor and the required steps within that range to calculate the exact number of possible readings that should be represented as binary bits. The only exceptions are the positioning readings (longitude and latitude) which were represented so as to reduce even more the bit representation required. In all cases linear quantisation is used. able I shows the lower and upper bound ranges for each sensor and the derived no of bits required representing each sensor reading. Ships latitude [18] is represented in degrees and tenths of a degree, measured in terms of degrees north or south of the equator. Latitudes are determined using standard shipboard methods i.e. a GPS receiver. Tenths are obtained by dividing the number of minutes by 6, and disregarding the remainder (Ignoring seconds). Coding is done with three digits; the first two digits are actual degrees, the last digit for tenths of a degree. Code 46° 41 as 466 (46° is coded as is, 41 divided by 6 is 6 5/6, 5/6 is disregarded); 33° 04 as 330 (33° is coded as is, 04 divided by 6 is 4/6 which is disregarded and coded as 0 in this case); 23° 00 as 230. Latitude can vary from 0° (coded 000) to 90° (coded 900). Quadrant of the globe (Qc) is used to specify whether the latitude is north or south. Ships Longitude [18] is also represented in degrees and tenths of a degree, measured in degrees east or west of the Greenwich Meridian. Values reverse at the international dateline. Tenths are obtained by dividing the number of minutes by 6, and disregarding the remainder (Ignoring seconds). Coding is done with four digits, with the leading (hundreds) figure coded as 0 or 1. The first three digits are actual degrees, the last digit for tenths of a degree. Code 142° 55 as 1429 (142° is coded as is, 55 divided by 6 is 9, the remainder is ignored); code 60° 31 as 0605 (60° is coded as 060, 31 divided by 6 is 5, the remainder is ignored); code 9° 40 as 0096 (9° is coded as 009, 40 is coded as 6); code 0° 16 as 0002 (0° is coded as 000, 16 is coded as 2). Longitude can vary from 0° (coded 0000 on the Greenwich Meridian) to 180° (coded 1800 on the dateline). Quadrant of the globe (Qc) is used to specify whether the longitude is east or west. Quadrant of the globe [18] varies according to your position with respect to the equator (0° latitude) and the Greenwich Meridian (0° longitude). If you are north of the equator (north latitude), Qc is coded as 1 when east of the Greenwich Meridian (east longitude), or as 7 when west of the Greenwich meridian; If you are south of the equator (south latitude), Qc is coded as 3 when east of the Greenwich meridian, or as 5 when west of the Greenwich meridian as shown in Figure(2). For positions on the equator, and on the Greenwich or 180th meridian, either of the two appropriate figures may be used.

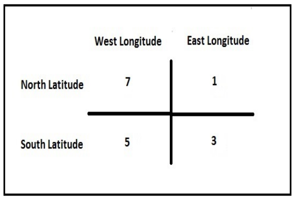


Fig. 2: Positioning according to quadrant of the globe

VIII. The proposed compression algorithm

The model of the proposed algorithm, the Average Marine Data Compression (AMDC) consists of three phases:

1. Quantizer: For our marine application, the data gathered from sensors are predictable, therefore it is essential to quantize the data to reduce the amount of bits needed to represent each reading in the binary representation.

2. Average Reading value (AR): It is calculated by summing the four readings after the current reading (Ri) from the sensors, then the deviation from the first reading is calculated as shown in equation 2 .

$$AR = R_i - \sum_{j=1}^{4} R_{i+j}. \qquad (2)$$

3. Arithmetic Coder: It calculates the arithmetic code for both (Ri) and (RA) values. After compression, the data is transmitted to the channel, Figure 3 shows this scheme.

TABLE I: Sensor reading ranges and the derived no of bits required to represent each reading

Parameters	**Lower value**	**Upper value**	**Quantized Representation**
Position			
Longitude	0	90	12
Latitude	0	180	13
Velocity			
Speed	0	76	7
Direction	0	8	3
Weather temperature	-50	50	8
Weather humidity	0	100	7
Wind direction	0	360	9
Wind speed	0	110	7
Water temperature	-2	36	6
Pressure barometric	800	1100	9
Salinity	0	44	7
Depth	0	10925	14
PH sensor	6.9	7.2	2

TABLE II: Sensor accuracy step level required

Sensor Measurement	Accuracy Step Level
Ship speed	0.5 km/h
Air temperature	0.5 deg. C
Air humidity	1.0 %
Wind Direction	1.0 deg
Wind speed	1.0 m/s
Water temperature	0.5 deg. C
Sea Level Pressure	0.5 mb
Salt level in water	0.5%
Depth of water	1.0 Meter
PH level	0.1

The Measured sensor readings are converted to a binary representation taking into account the quantization of each sensor reading as shown in Table I. Quantization readings are represented by N bits in an analogue to digital converter (ADC), where N is the resolution of the ADC .

IX. Experiment

In this paper, we propose a specific data formatting for the data gathering application and compress this data to reduce the size of data transmission for sensor nodes over the marine network channel. However, in our proposed MANET over VHF radio frequencies, the transmission bandwidth used is 9.6 Kb/Sec. By reducing data size less bandwidth is required for sending and receiving data. The data compression is one effective method to utilize limited resources of WSNs, therefore its crucial to compress the data before sending over the transmission media. We have simulated a lossless data compression algorithm particularly suitable for the limited storage and computational resources of a wireless sensor network node. We have simulated the algorithm and used it on our marine data that was obtained from an AIS live system [19], [20]. Note some of the readings were obtained using interpolation. we compare between our proposed algorithm and the Arithmetic coding algorithm and evaluate the performance of the algorithm using the compression ratio metric for the compressed data at the originator node. We obtain the compression ratio 90.11 % and 89.25 % for the two data samples respectively.

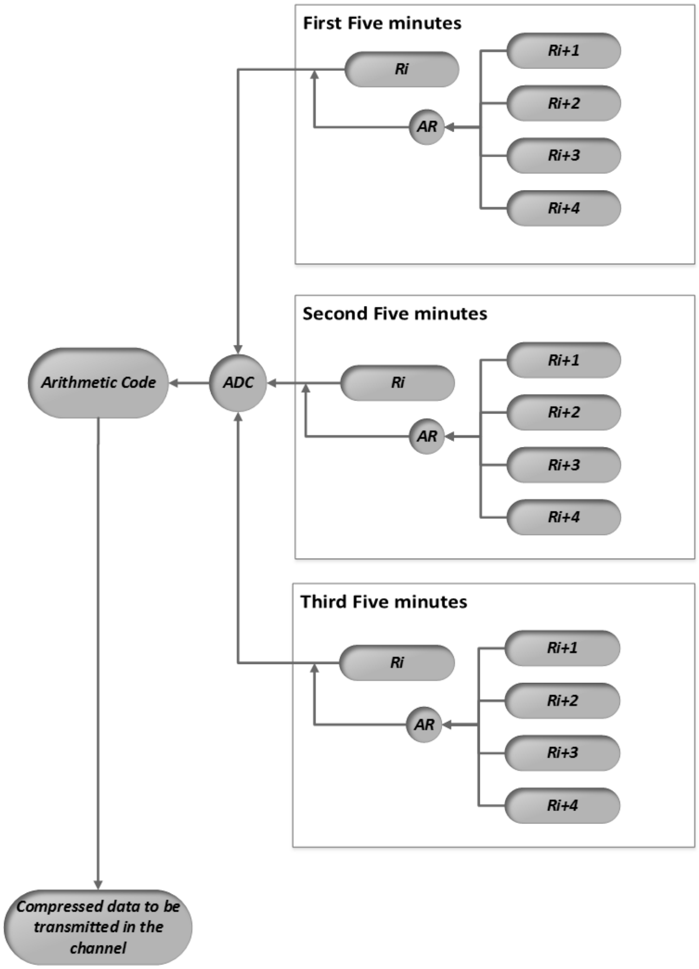


Fig. 3: AMDC Proposed Model

X. Result and comparison

The scheme presented can be implemented on sensors in a WSN. In our application we used 11 sensors, which sense values once each minute. According to our quantization method in Table I, we have 104 samples for one reading for the whole 11 sensors. The performances of the schemes were analysed according to the number of bits required to transmit the acquired data and the compression ratio. During simulation, attention was focused only on the bits required to compress the data. Sets of data were considered, representing the (Position, velocity, humidity, temperature, wind speed, wind direction, barometric pressure, salinity, depth and PH) values collected during 15 minutes in the marine environment. Considering the acquisition time of 15 minutes, the sensor should acquire one values for each minute. For 11 sensors we have 11 readings, the total transmitted bits for the compressed data is 2085, while the total amount of compressed values is of 206 bits. For the two examples discussed (data1, data2) the results can be seen in Figure 4.

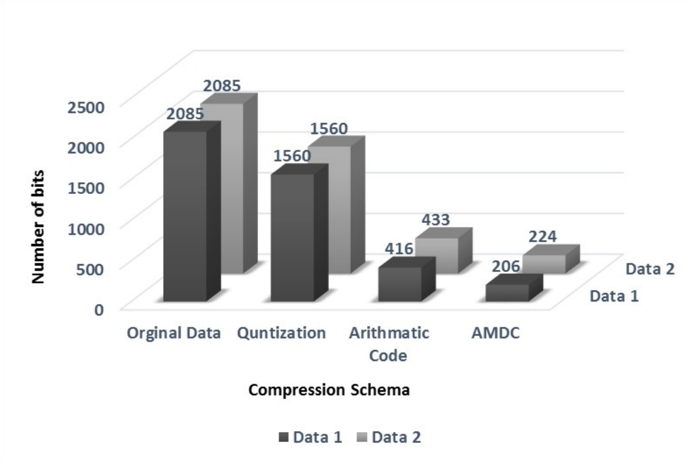


Fig. 4: comparison between Arithmetic code and AMDC

The metric used to compute the performance of the data compression algorithm is the compression ratio and is defined as the ratio between the size of the compressed file and the size of the source file as shown in Equation 3 :

$$CR = 100 * (1 - C_{Size}/O_{Size}) \tag{3}$$

where CR (Compression ratio) , C_{Size}(Compression size) and O_{Size} (Orginal size) respectively, the size of the compressed and the uncompressed bit stream.

Table III summarizes the results obtained by applying the proposed compression algorithm in contrast with applying arithmetic code compression for two data sets that represent two different input streams of marine sensor data. The table shows clearly that the proposed (AMDC) algorithm outperforms arithmetic code in compression rate for both data sets applied.

TABLE III: Comparison Ratio for Data1 and Data2

	Arithmetic code compression	**Proposed AMDC compression**
Data 1	80.04%	90.11%
Data 2	79.23%	89.25%

XI. Conclusions

MANET networks in the marine environment using VHF technology available on the majority of ships and vessels in order to gather important sensor data is a promising research field to overcome the high cost burden of satellite communications currently in place. But on the other hand due to bandwidth limitations of the VHF channel, minimizing transmission data redundancy overhead is essential for efficient use of the transmission channel. For our marine application, the predictability of gathered sensor data makes it beneficial to quantize the data to reduce the amount of bits needed to represent each reading in the binary representation. Applying this quantization in conjunction with the proposed compression algorithm (AMDC) has proved affective data compression rates in comparison with the major known compression methods.

References

[1] N. Kimura and S. Latifi, “A survey on data compression in wireless sensor networks,” in *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, vol. 2. IEEE, 2005, pp. 8–13.

[2] J. G. Kolo, S. A. Shanmugam, D. W. G. Lim, L.-M. Ang, and K. P. Seng, “An adaptive lossless data compression scheme for wireless sensor networks,” *Journal of Sensors*, vol. 2012, 2012.

[3] R. Mohsin and J. Woods, “Performance evaluation of manet routing protocols in a maritime environment,” in *Computer Science and Electronic Engineering Conference (CEEC), 2014 6th*. IEEE, 2014, pp. 1–5.

[4] A. K. Maurya, D. Singh, and A. K. Sarje, “Median predictor based data compression algorithm for wireless sensor network,” *International Journal of Smart Sensors and Ad Hoc Networks*, vol. 1, no. 1, pp. 62–65, 2011.

[5] C. M. Sadler and M. Martonosi, “Data compression algorithms for energy-constrained devices in delay tolerant networks,” in *Proceedings of the 4th international conference on Embedded networked sensor systems*. ACM, 2006, pp. 265–278.

[6] T. Bordley, “Linear predictive coding of marine seismic data,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 31, no. 4, pp. 828–835, 1983.

[7] M. Schmalz, G. Ritter, and F. Caimi, “Data compression techniques for underwater imagery,” in *OCEANS’96. MTS/IEEE. Prospects for the 21st Century. Conference Proceedings*, vol. 2. IEEE, 1996, pp. 929–936.

[8] D. F. Hoag, V. K. Ingle, and R. J. Gaudette, “Low-bit-rate coding of underwater video using wavelet-based compression algorithms,” *Oceanic Engineering, IEEE Journal of*, vol. 22, no. 2, pp. 393–400, 1997.

[9] “Digital signals - sampling and quantization.” [Online]. Available: http://www.rs-met.com/documents/tutorials/DigitalSignals.pdf

[10] H. Johansson, “Sampling and quantization,” *Academic Press Library in Signal Processing: Signal Processing Theory and Machine Learning*, vol. 1, p. 169, 2013.

[11] W. Sheppard, “On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale,” *Proceedings of the London Mathematical Society*, vol. 1, no. 1, pp. 353–380, 1897.

[12] R. M. Gray and D. L. Neuhoff, “Quantization,” *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2325–2383, 1998.

[13] A. Kondoz, “Sampling and quantization,” *Digital Speech: Coding for Low Bit Rate Communication Systems, Second Edition*, pp. 23–55.

[14] T. Anttalainen, *Introduction to telecommunications network engineering*. Artech House, 2003.

[15] M. Mahoney, “Data compression programs,” 2008.

[16] S. Kodituwakku and U. Amarasinghe, “Comparison of lossless data compression algorithms for text data,” *Indian journal of computer science and engineering*, vol. 1, no. 4, pp. 416–425, 2010.

[17] D. P. Stern, “Latitude and longitude,” *Web page, NASA, Goddard Space Flight Center, Greenbelt, Maryland*, vol. 17, 2004.

[18] U. D. O. COMMERCE.

[19] F. Inc., “Live marine traffic.” [Online]. Available: https://www.marinetraffic.com

[20] "Tonbridge-weather." [Online]. Available: http://www.tonbridge-weather.org.uk/home.htm

An Adaptive Learning Mechanism for Selection of Increasingly More Complex Systems

Fouad Khan
Department of Environmental Science and Policy
Central European University
Hungary, Budapest

***Abstract*— Recently it has been demonstrated that causal entropic forces can lead to the emergence of complex phenomena associated with human cognitive niche such as tool use and social cooperation. Here I show that even more fundamental traits associated with human cognition such as 'self-awareness' can easily be demonstrated to be arising out of merely a selection for 'better regulators'; i.e. systems which respond comparatively better to threats to their existence which are internal to themselves. A simple model demonstrates how indeed the average self-awareness for a universe of systems continues to rise as less self-aware systems are eliminated. The model also demonstrates however that the maximum attainable self-awareness for any system is limited by the plasticity and energy availability for that typology of systems. I argue that this rise in self-awareness may be the reason why systems tend towards greater complexity.**

Keywords*— Adaptive Learning, Complexity, Self-awareness, Good regulator theorem, Adaptive Selection

I. Introduction

One of the by-products of the revolution in information technology over the last three decades has been our enhanced capacity to visualize, model and understand complex phenomena. This has allowed us to identify and visualize key traits associated with complexity such as self-similarity [1] and recursion [2], interconnectedness of elements [3], high sensitivity to initial conditions [4], and theorize about the sources of these traits [5-9] and evolution of complex systems [10]. These developments though have not brought us much closer to eliminating widespread skepticism about either our ability to build predictive models of complex phenomena [11] or arrive at feasible mechanisms to describe the emergence and selection of such phenomena associated with complexity as human cognition [12], though some of the findings are already being incorporated in systems analysis, design and architecting [13]. It has also been shown that in clustering systems without noise reaching consensus is directly proportional to the size of group [14].

Recently however, it was demonstrated that traits associated with the human cognitive niche such as tool use and social cooperation can naturally emerge under the action of causal entropic forces [9]. Here, through a simple model, I demonstrate that even more rudimentary complex phenomena associated with human cognition such as 'self-awareness', can naturally emerge in systems in response to 'internal stimuli' as these internal stimuli eliminate less 'self-aware' systems.

Mechanisms proposed so far only look at external stimuli (for instance in the case of natural selection) for evolution of complexity. The mechanism proposed here acknowledges that drivers of evolution of complexity can be transformations internal to the system as well.

This paper presents a model that shows how internal stimuli through a proposed new mechanism leads to the selection of ever more complex systems.

The work presented here can be seen as a corollary of the good regulator theorem [15] and has been done to show the limitations other works that propose entropic measures as drivers for complexity [9] in a competitive environment but ignore internal stimuli; in the presence of which, competitive environment is not necessary for evolution of complexity.

II. Materials and Methodology

To construct the model we start with a system which is a 'good regulator' of itself [15]. It has been shown that any good regulator of a system is also a model of the system [15]. So if R is a good regulator of System S, then it is both a) internal to the system and b) a model of the system. Also for every 'real world' state the system S assumes, R (being a model of S) assumes a corresponding 'model' state. For the purposes of development of this model 'self awareness' (to be denoted by Δ) now is defined as the change in internal model R with change in system S.

$$\Delta = \frac{dR}{dS} \quad (1)$$

Defined in this manner, self-awareness stops being a binary property but instead can be represented by a continuous bounded function (with values between 0 and 1). Instead of just either having or not having 'self-awareness', systems can have varying degrees of self-awareness; self-similarity for instance being one of the cruder forms (lower degree) of self-awareness. Every system can be imagined to have an internal model of itself within it, the question remains only of quantifying the degree of accuracy of that model.

Imagine now that starting from a state S_o, our system goes to a critical state S_c at which the system ceases to exist due to internal stimuli. At state S_o, the internal model of the system is in state R_o. However, the internal model (which is also a good regulator) also has a state R_c at which the system realizes the threat posed by the internal stimuli and adjusts its state before it reaches the critical state S_c. Any system for which the time T_R taken for R to reach R_c is smaller than the time T_S taken for S to reach S_c would have a longer time of existence compared to a system where $T_S<T_R$. This is the survival advantage that systems with higher Δ would have, given all else is equal. So, for a regulator to be good enough to provide survival advantage;

$$T_R<T_S$$

Where;

$$T_S = \frac{S_c - S_o}{\frac{dS}{dt}} \quad (2)$$

And

$$T_R = \frac{R_c - R_o}{\frac{dR}{dt}} \quad (3)$$

Substituting in equation 1, for an internal model to be good enough to provide survival advantage;

$$\frac{R_c - R_o}{\frac{dR}{dt}} < \frac{S_c - S_o}{\frac{dS}{dt}} \quad (4)$$

Given that dR = ΔdS;

$$\frac{R_c - R_o}{S_c - S_o} < \Delta \; (5)$$

The probability of condition specified in equation 5 being true increases with increasing Δ (where Δ is some function of

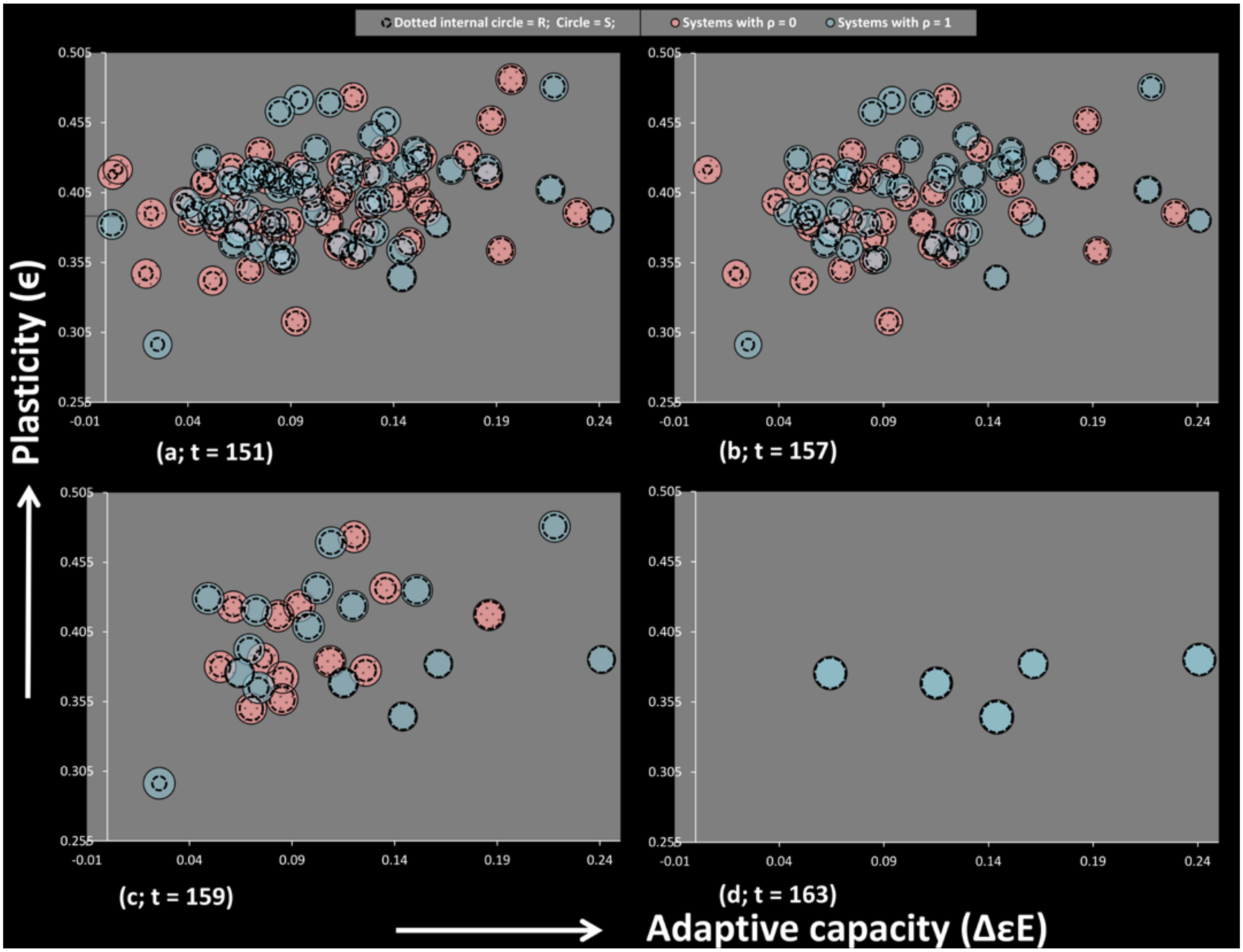


Fig. 1. Systems with lower adaptive capacity (ΔεE) die-off under adaptive selection as universe evolves over time-steps a) 151, b) 157, c) 159, d) 163. Bubbles with dotted fill are systems with agency (ρ) = 0, while bubbles with solid fill are systems with agency (ρ) = 1. Bubble size indicates value of one system state variable X. Size of the dotted outlined bubble inside bigger bubbles indicates internal model value x for the same variable X in the internal model R. As can be seen in d at time-step 163, the surviving systems are ones with very high self-awareness (dotted outline is closest to solid outline)

the internal state variable/s of S with a range between 0 and 1) or 'self-awareness'. What this results seems to imply is that not only is a good regulator one which is a model of the system being regulated, but the better this internal model of the system is -or the higher the self-awareness of the system- the more probable it is to survive (in response to internal threats to its existence).

A simple numerical model consisting of a universe with hundred systems of varying self-awareness was built to further demonstrate how this mechanism naturally selects for systems with higher self-awareness. A binary property ρ to be called 'agency' was also introduced in the model. When R equaled R_c for any system, the system readjusted only if ρ equaled 1. Overtime, we expected to see more systems with the agency switch 'on' (ρ = 1) survive as opposed to those where ρ was equal to 0. The magnitude of the readjustment depended upon the 'plasticity' of the system. Plasticity was defined as the deformation in S, per unit of available energy E, normalized to the initial value of S. Plasticity, denoted by ϵ can be expressed as;

$$\epsilon = \frac{dS}{ES} \ (6)$$

Further, R_c depended on how quickly the system was able to identify the need for a readjustment. This property was termed 'agility'; defined as the difference between the system critical value (S_c) and internal model critical value (R_c), normalized to the system critical value S_c. Agility, denoted by τ can be expressed as;

$$\tau = \frac{(S_c - R_c)}{S_c} \ (7)$$

Four parameters are monitored across the set of 'living' systems as our universe evolved and some systems were eliminated due to S having reached critical value S_c; i) the average self-awareness Δ_{ave}; ii) ratio of number of systems with 0 agency against number of systems with agency equal to 1, ρ_R; iii) average agility τ_{ave} and iv) average plasticity ϵ_{ave}.

III. Results and Discussion

One immediately observable fact was that all these properties across the universe evolved in bursts (spasmodically) in a manner reminiscent of scale-free networks [3].

Average self-awareness for the set of living systems was indeed seen to increase with elimination of less self-aware systems, though it was observed that the maximum attainable self-awareness for any system was limited by the product of self-awareness, plasticity and energy for that system typology.

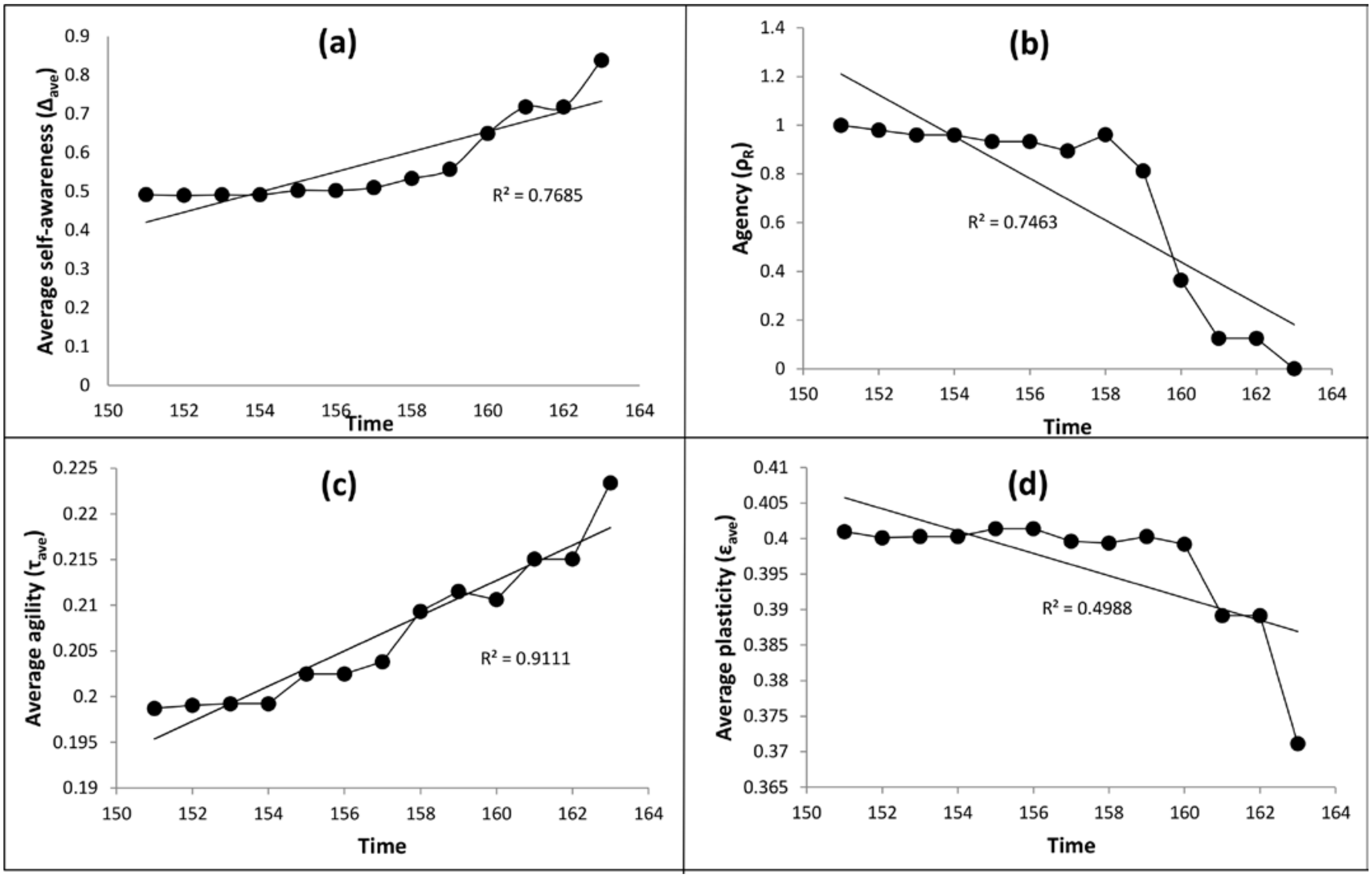


Fig. 2. Average Self-awareness of the set of living systems increases over time; b) Non-reactive systems die-off as the ratio of non-reactive to reactive systems decreases over time; c) Average agility of the set of living systems increases over time; d) Average plasticity of the set of living systems decreases over time

We term this product the adaptive capacity. Figure 1 shows the elimination process at four time steps during the model run.

Figure 2 shows how the monitored properties evolved over time for the universe of living systems with average self-awareness and agility increasing and ratio of positive agency over null agency systems decreasing as expected, and the average plasticity decreasing. The rise in plasticity is somewhat surprising. One should expect that the more plastic a system is, the more adaptable it should be, and hence the more resilient. What we see instead is that the systems that survive are the ones with lower plasticity.

However, from equations 1 and 6 we deduce that the change in model normalized to the original system state is equal to the product of self-awareness, plasticity and energy availability.

$$\frac{dR}{S} = \Delta\epsilon E \quad (8)$$

From equation 8 we can see that plasticity (ϵ) and self-awareness (Δ) are inversely related. Upon consideration this result does appear to make intuitive sense. Plasticity is a measure of how much change R can incur in S, while self-awareness is a measure of how R changes with changes in S. For any given system, the internal model can be made of either energy or matter, however in most cases, the internal model substitutes information for what is material in a system; actual quantities are replaced by say, a number representing that quantity. A state variable in the internal model say R though is more likely to either be ‘information’ or energy, while S, the corresponding system state variable, can be expected to have more of a material component. Imagine for instance a refrigerator, say S to a model of the refrigerator as it exists in your mind, say R. The former has a lot more material content compared to the latter. Self-awareness thus can be conceptualized as the amount of change incurred in informational content with change in real world material counterpart. Plasticity then is a measure of how that change in information comes back and affects a change in its real world material counterpart. This loop –system affecting model affecting system- is the essence of sentience and consciousness. The term ΔϵE arrived at in equation 8 defines the upper bounds for this property for any given system. For any given system ‘typology’ (all systems with the same plasticity and energy availability), the product ϵE determines the upper bounds of adaptive capacity.

IV. Conclusions

This model demonstrates not only how systems naturally tend towards greater self-awareness but also how the potential for self-awareness is restricted by the plasticity of the system and the energy availability. For any given typology (here defined by the product of plasticity and energy) thus, we will see more self-aware systems survive over longer runs, but no system can rise above the limitations imposed upon it by its typology. For planetary systems for instance, the energy available as electromagnetic forces is very weak as electromagnetic forces are weak at that scale. Energy available as gravitational force, though stronger is still comparatively weaker in terms of its ability to cause strain in the system (hence lower plasticity). This means that ΔϵE has a low value compared to organic systems where electromagnetic forces act on organic matter (much more malleable hence susceptible to higher strain and having higher plasticity). Since both ϵ and E are quantifiable terms, establishing indicative values of ϵE for different system typologies should be trivial. It could be easy to show why the organic brain with its high material malleability and energy availability offers such a generous nursery for the rise of self-awareness.

It should also be noted that for self-awareness Δ to be higher, the variables that define the state of internal model R should have higher number of stronger correlations with their corresponding counterparts in system S; the variables that define the state of system S. Higher self-awareness thus is a measure of higher number of stronger correlations between internal state variables of a system. This implies greater internal interconnectivity and thus greater complexity within the system. This means that the mechanism proposed here –an adaptive selection of better regulators- also elaborates how systems naturally tend towards higher complexity.

Since like the good regulator theorem this work is applicable to all systems from ‘a cow’s digestive system’ [15] to national politics, examples of the mechanism proposed here can be seen in the process of regulation in many complex systems such as cities and national economies where increasing disparity and difficulty to model, increases the energy cost of regulation.

In future the research shall be expanded by empirical analysis of regulation data from complex systems such as cities and national economies.

References

[1] B. Mandelbrot, "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," *Science,* vol. 156, pp. 636-638, May 5, 1967 1967.

[2] D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*: Basic Books, Inc., 1979.

[3] A.-L. Barabási, *et al.*, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications,* vol. 272, pp. 173-187, 1999.

[4] S. Wolfram, *A new kind of science*, 2001.

[5] L. M. A. Bettencourt, "The Origins of Scaling in Cities," *Science,* vol. 340, pp. 1438-1441, June 21, 2013 2013.

[6] B. B. Mandelbrot, *The fractal geometry of nature*: W. H. Freeman, 1983.

[7] [G. B. West and J. H. Brown, "A general model for the origin of allometric scaling laws in biology," *Science,* vol. 276, p. 122, 1997.

[8] G. B. West, *et al.*, "The Fourth Dimension of Life: Fractal Geometry and Allometric Scaling of Organisms," *Science,* vol. 284, pp. 1677-1679, June 4, 1999 1999.

[9] A. D. Wissner-Gross and C. E. Freer, "Causal Entropic Forces," *Physical Review Letters*, vol. 110, p. 168702, 2013.

[10] E. Chaisson, *Cosmic Evolution: The Rise of Complexity in Nature*. Cambridge, MA: Harvard UP, 2001.

[11] N. Taleb, *The black swan: the impact of the highly improbable*: Penguin, 2008.

[12] T. Nagel, *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False.* New York: Oxford University Press, 2012.

[13] C. H. Dagli, *et al.*, *Intelligent Engineering Systems Through Artificial Neural Networks: Computational Intelligence in Architecting Complex Engineering Systems*: American Society of Mechanical Engineers, 2009.

[14] Shang, Y. and R. Bouffanais, *Influence of the number of topologically interacting neighbors on swarm dynamics.* Scientific Reports, 2014. **4**: p. 4184.

[15] R. C. Conant and W. Ross Ashby, "Every good regulator of a system must be a model of that system †," *International Journal of Systems Science*, vol. 1, pp. 89-97, 1970/10/01 1970.

Classification model of arousal and valence mental states by EEG signals analysis and Brodmann correlations

Adrian Rodriguez Aguiñaga and Miguel Angel López Ramírez
Instituto Tecnológico de Tijuana
Calzada del Tecnolo´gico S/N, Toma´s Aquino, 22414
Tijuana, B.C. Me´xico

María del Rosario Baltazar Flores
Instituto Tecnológico de León
Av. Tecnológico S/N
Industrial Julia´n de Obrego´n, 37290
Leo´n, Gto. Me´xico

***Abstract*—This paper proposes a methodology to perform emotional states classification by the analysis of EEG signals, wavelet decomposition and an electrode discrimination process, that associates electrodes of a 10/20 model to Brodmann regions and reduce computational burden. The classification process were performed by a Support Vector Machines Classification process, achieving a 81.46 percent of classification rate for a multi-class problem and the emotions modeling are based in an adjusted space from the Russell Arousal Valence Space and the Geneva model.**

***Keywords*—Emotions; Affective Computing; EEG; SVM; Wavelets;**

I. Introduction

The development of technologies that allow interaction between a user and a computer in a more natural way, has been one of biggest challenges in recent decades.

Since Rosalind Picard founded the Affective Computing Group at MIT and propose theories to establish a better understanding of the impact of technology on the emotional states [1], a wide range of important developments focused in the human-machine interaction have been developed; being one of the most relevant the analysis of emotional states, due the great importance of emotions in our daily communication.

To date many techniques to analyzing the physiological expressions of an emotion have been developed; however, most of them are susceptible to be manipulated by users and provide unreliable information. One of the main proposals to resolve this problems, are the analysis of bio-medical signals (biosignals), such as heart rate, breathing rate and the behavior of neural signals that properly processed can be an important and reliable information source [2]; being the brain signal analysis, one of the techniques that has gained an increased demand in the past decades, due the recent developments in Brain Computer Interfaces (BCI), signal processing and pattern recognition algorithms facilitates the analysis, development and implementation of affective technologies. Also previous efforts to perform emotional states recognition, has been reached up to 80% of classification rates (Table I), promising results if we consider that a person can recognize an emotional state from another only the 88% of the time.

This paper presents a strategy for recognition and classification of emotional states through analysis of electroencephalography (EEG) signals and a data reduction strategy based on the correlation between electrodes and a bounded region delimited by a Brodmann model analysis.

A. Related work

Recent advances in the analysis of biological signals achieves promising results as presented in Table I.

TABLE I: Previous emotion recognition developments and it recognition rate.

Technique	Recognition Rate %	Reference
MPC	64	[3]
MDC	74.11	[4]
SVM	66.7	[5]
SVM	93.5	[6]
SVM	77.8	[7]
KNN	82.27	[8]
NN	43.14	[9]
NN	60	[10]
NN	93.3	[11]

Many of this research's achieves more than 88% of recognition rate, however, by performing binary classifications task or employing physiological characteristics as reference, since the associated to biological signals analysis shows a significant decrements in the recognition rate. Although a wide variety of techniques to perform classification are being implement to date, SVM and NN has shown the best performance.

Section II, describes a proposed emotion characterization model to perform a multidimensional classification process; Section III, describes a proposed strategy to reduce data in an EEG analysis by establishing a bounded regions of an electrode correlations to Brodmann regions and the features extraction and arrangement; Section IV, describes the configuration for the classification process and the implemented methodology to perform the experimentation processes; Section V, presents the results obtained by implementing the proposed strategies and Section VI, discusses the conclusions of this work.

II. EMOTIONS

Define a concept of emotion, could be a notorious problem without a proper review in the psychology advances.

Since we are proposing a classification task which involves a systematic process, the adopted definition of emotion has to satisfy this constraint. Klaus R. Scherer, define emotions as *an episode of interrelated, synchronized changes in the states of all or most of the five*[1] *organismic subsystems* [2] *in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism* [12].

However, the high diversity of theories associated to emotional states, implies that a very extensive endeavors must be performed just to pose an appropriate model to perform an adequate classification space. Fortunately Paul Ekman, Wallace Friesen, James Russell and Klaus R. Scherer propose theories that could be used to overcome most of this issues.

Ekman and Friesen propose the big six and the micro-expressions theory which generalize emotional gestures and provide tags to emotional states [13], Russell proposed an Arousal and Valence Space (AVS), where an emotional state could be represented as combination levels [14] and the Geneva model to provide a higher abstraction level by the addition of a dominance parameter to the AVS [12]. The classification classes are then conformed by referencing each experimental case to a certain emotion, with the following parameters:

- Arousal: ranges from inactive (e.g. uninterested, bored) to active (e.g. alert, excited).
- Valence: ranges from unpleasant (e.g. sad, stressed) to pleasant (e.g. happy, elated).
- Dominance: ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything).

The proposed space [3] are configured as in Figure 1, to establish a four class problems defined as high arousal and high valence ([HA-HV] Happiness-Elation), high arousal and low valence ([HA-LV] Relax-Calm), low arousal and high valence ([LA-HV] Boredom-Sadness) and low arousal and low valence ([LA-LV] Hostility-Anger), combinations.

A. Data characterization

The lack of benchmarks databases, is another important challenge in the emotional states classification process due that standardized databases as the International Affective Picture System Digitized (IAPS) and the International Affective Digital Sounds (IADS) [15] [16], can only be employed to perform experimental setups and databases as the bu-3DFE, PhysioNet, iBUG project and DEAP dataset, which provides an experimental generalization due it are conformed from biological signals of a specific group of participants are not standardized.

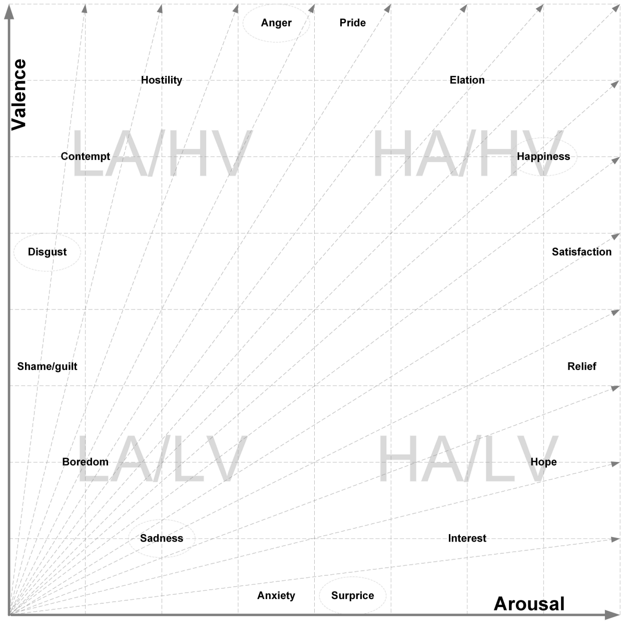


Fig. 1: AVS model: The emotion distribution provided by the Russell AVS and the discrete tags by the Ekman model.

- bu-3DFE [17]: a database by GAIC lab of is a collection of facial expressions records.
- PhysioNet [18]: a very wide a collection of bio signals provided by the National Institute of Biomedical Imaging and Bioengineering.
- The iBUG project [19]: a set of databases of the human behavior analysis through a bio signals characterization recompiled by the Intelligent Behavior Understanding Group.
- Database for Emotion Analysis using Physiological Signals (DEAP)[20]: is a collection of biomedical information from thirty two participants submitted to emotional stimulus.

To perform our classification task, we implement the DEAP dataset, which is up to our best knowledge the most complete database and are developed under the specific purpose of record the psychological responses of an emotion in arousal and valence levels. Also DEAP is already related to several other research projects [21] [22] [23] [24].

III. SIGNAL CONDITIONING

The large amounts of data associated to an EEG classification process are also an important issue to be addressed and to handle whit it, a strategy that reduces the processed data by implementing an electrode discrimination process and a reduction of the spectral information by a frequency bands cutoff, are presented.

[1] Information processing (CNS), Support (CNS, NES, ANS), Executive (CNS), Action (SNS), Monitor (CNS.)

[2] (CNS), central nervous system; (NES), neuro-endocrine system; (ANS), autonomic nervous system; (SNS), somatic nervous system. The organismic subsystems are theoretically postulated functional units or networks.

[3] This model do not consider the dominance parameter to keep the experiment in an four class problem.

A. Brodmann analysis

Several researches has sought the establishment of a region based criteria which relates emotional activity to brain regions, an many of them suggest an association between the occipital and temporal cortex regions with emotional processes [4].

The proposed strategy suggested in this work, establish a relationship between cortex regions an its relevancy in the classification (TableII).

Figure 2.

TABLE II: Boadmann association regions

Process	Brodmann regions	Cortex regions
Visual	17,18,19,20,21,37.	Temporal lobe Occipital lobe
Auditory	22,41,42.	Temporal lobe
Sensitive	1,2,3,4,5,7,22,37,39,40	Parietal lobe
Motor	4,6,44,9,10,11,45,46,47	Temporal lobe Frontal lobe

Also Table III, presents the association between electrodes to the associated regions an its correlation to conform a region based model, presented in Figure 2.

TABLE III: Brodmann regions and it associated electrodes

Associated region	Electrode
Frontal Temporal 7	F7
Frontal cortex 5	FC5
Parieto Temporal 7	T7
Parietal cortex 5	CP5
Parieto Occipital 7	P7
Parieto Occipital 3	P3
Occipital 1	O1
Occipital 2	O2
Parieto Occipital 4	P4
Parieto Occipital 8	P8
Frontal cortex 6	C6
Temporal 8	T8
Frontal cortex 5	C5
Frontal Temporal 8	F8
Parietal cortex 6	CP6

This strategy only considers only fifteen electrodes from a differential 10/20 model and achieves a reduction of 38.5% of the total processed data.

B. Filtering

Noise and crosstalking effects are another important problem to be addressed to identify a specific behavior EEG signals, an a wide range of signal processing techniques must be implemented just to handle with this problem [2][25]. EEG noise are generally related to instrumentation and physiological activity as the vascular, muscular or ocular movements and to attenuate this, a derivative of a surface LaPlacean filter (SL) suggested by M. Murugappan were implemented to lead a proper filtering process [26] [27]. This technique filters out the signals originated outside of the skull and emphasizing the electrical activity that occurs near an electrode by attenuating EEG activity and improving the spatial resolution of the signals, when it became common to all electrodes.

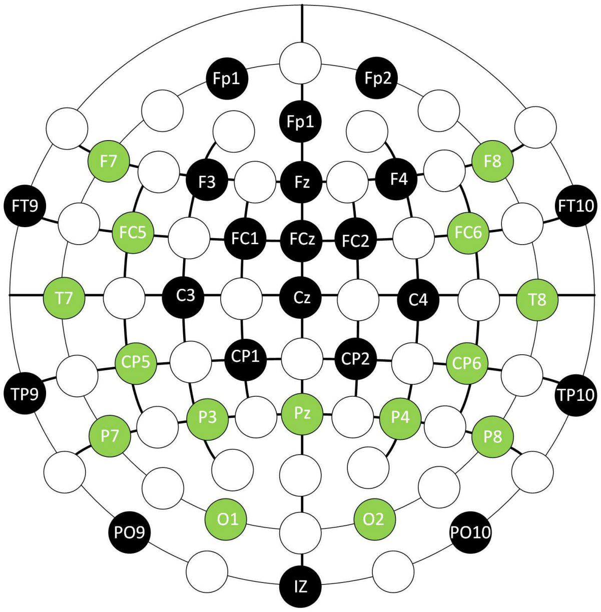


Fig. 2: Selected electrodes as primary electrodes, this association are modeled with the 10/20 Biosemis system as reference.

$$X_{new} = X(t) - 1/N_E \sum_{i=1}^{N_E} X_i(t), \qquad (1)$$

where X_{new} is the filtered signal, $X(t)$ the raw signals and N_E the number of neighbor electrodes.

Also a Independent Component Analysis Blind Source Separation technique (ICA-BSS) were implemented to include the uncorrelated information associated by the electrode discriminant process. This process also reduces redundancy and preserve references of the uncorrelated electrodes, by the decomposition of the signal into its constituent independent components X_{new}, considering a multi-channel signal problem defined as $y(n)$ and the constituent components of the signal as $y_i(n)$, then

$$P_Y(y(n)) = \prod_{m}^{i=1} p_y(y_i(n)) \ \ \forall n, \qquad (2)$$

where P_Y is the probability distribution set, $p_y(y_i(n))$ are the marginal distributions and m is the number of independent components, performed over all the of electrodes (Figure 3).

To consider all the mixing structure process are necessary to estimates all the independent sources from the signal and for most cases this information are unavailable, and the assumption that this mixing process occurs as an instant case are widely accepted in the analysis of biological signals such as EEG where the signals are composed by a narrow bandwidth and a low sampling frequency[5], the formulating BSS model are modeled as

[4] Encompassing the hypothalamus, pituitary gland, amygdala and hippocampus.

[5] This case assumes that signals reach the sensors at the same time.

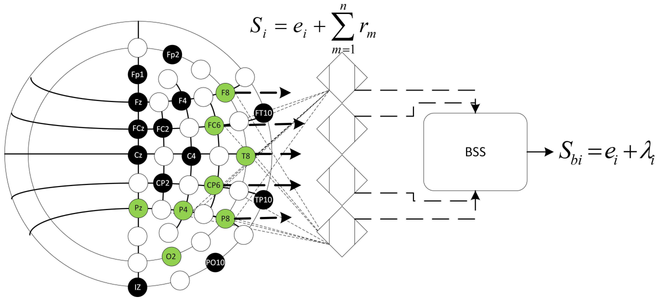


Fig. 3: S_i are the overlay sensor signals, e_i are the electrodes references and $\sum_{m=1}^{n} r_m$ is the attached signal of the electrodes. S_{bi} is the reference signal of a electrode after the filtering process, where e_i is the electrode signal of a single electrode plus the λ_i remaining overlay signals of the non-primary electrodes. The main approach is that the reference energy at the focal points of the sensor will be increased, so that by using a process of BSS, may obtain independent vectors corresponding to each region.

$$z = H_s(n) + v(n), \tag{3}$$

H are the mixing matrix, $v(n)$ denotes the signal source vectors and z is a mxn matrix from n samples, from a m-dimensional array of electrodes activity.

C. Rhythms

The frequency bands associated to brain activity are denominated as rhythms (Table IV), and several studies suggest that are associated to a specific mental tasks; as another measure to perform signal conditioning a band pass filter of 0.5 Hz to 47 Hz, were implemented to consider only the frequencies of rhythms [25] [28].

TABLE IV: Brain rhythms [25].

Rhythms	Frequency band (Hz)
Delta	0.5 to 4
Theta	4 to 8
Alpha	8 to 12
Mu	8 to 13
Beta	12 to 30
Gamma	> 30

D. Feature extraction

1) Wavelet decomposition: Despite the fact that to date, many techniques to perform EEG feature extraction have been developed; Wavelet Transform (WT), still are one of the most reliable feature extraction methodology in the EEG analysis [29] [27]; one of the most important aspects of the WT, is that transformation coefficients can be employed directly as features for the classification problem and unlike the Fourier transform it provides a commitment between the spatial and temporal information of the signal[6], which are very useful in the analysis of biological signals.

Also WT handles with the non-stationary nature of the EEG signals by expanding, contracting and shifting a function $\psi_{a,b}$ called "*mother*" wavelet, defined by as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \quad a, b \in R, a > 0, \tag{4}$$

R are in a wavelet space and a is a scaling factor, b is the shifting factor.

Due EEG signals are non-stationary and their local waves are often spatially variant to decompose a signal into the elementary building blocks. WT allows a well localized in time and frequency due small scales reflects the high frequency components of the signal and large scales reflects the low frequency components of the signal, allowing the representation of the temporal features of a signal at different resolution [30].

The selection of an appropriate wavelet for a given signal, still implies an extensive search; to perform the presented feature extraction Daubechie 6 (DB6), were selected due it shows the best performance in a heuristic experimental task and also satisfy the following admissibility condition to handle whit the non-stationary propriety of EEG

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty, \tag{5}$$

where $\Psi(\omega)$, is the Fourier transform of $\psi_{a,b}(t)$.

2) Approximation and detail based features: The WT describes signal in terms of coefficients by representing their energy content in a specified frequency region and this coefficients conforms the classification arrangements.

This signal $f(t)$ can be decomposed as

$$f(t) = \sum_j \sum_k d_{j,k}\psi_{j,k}(t) = \sum_j f_j(t), \tag{6}$$

where $j,k \in Z$ and $\psi(t)$ is a mother Wavelet and the coefficients $d_{j,k}$ is the inner product

$$d_{j,k} = \langle f(t), \psi_{j,k}(t)\rangle = \frac{1}{\sqrt{2^j}} \int f(t)\psi(2^{-j}t - k)dt, \tag{7}$$

From the wavelet coefficients $d_{j,k}$, the energy of the details of f at level j can be expressed as

$$E_j = \sum_k d_{j,k}^2. \tag{8}$$

If the total energy of the details is denoted as $E_t = \sum_j E_j$, then the percentile energy corresponding at level j is:

$$\varepsilon_j = \frac{E_j}{E_t} \times 100, \tag{9}$$

The level j is associated with frequency band, ΔF, given by:

$$2^{-j-1}F_s \leqslant \Delta F \leqslant 2^{-j}F_s, \tag{10}$$

where F_s is the sampling frequency.

[6] The time-frequency representation are performed by a set of filters, that limits the frequency domain by half and decompose the signal into approximation coefficients (AC) and detail coefficients (DC) and this process are performed by iterative and complementary filters

IV. Classification

A. *Experimental setup*

Figure 4, presents the four classes distribution generated by a statistic distribution of from the evoked potential to describes each experiment case. The classes are conformed by sets of ten experimental setups.

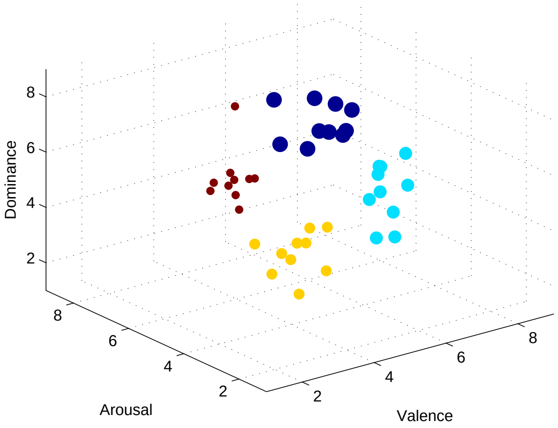


Fig. 4: Evoked potentials distribution arrangement base on the $\bar{x}$ EP of the thirty two participants.

1) Cases: A sub-cases configuration were proposed to evaluated the performance of this classification strategy; under distinct information configurations, conformed by an arrangements selection of the three most representative experimental cases, as presented in Figure 5.

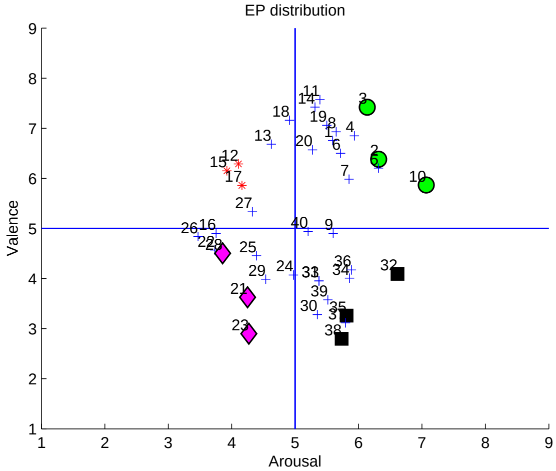


Fig. 5: Most representative experimental cases associated to each class, selected by a Mahalanobis distance.

This configuration shows the performance obtained under distinct stimulus over all the participants, and to ensure that each of this arrangements contains the information of different experimentation trials constructed under the following configuration:

- Case A: Information of a the experimental cases 3,12,28 and 32.
- Case B: Information of a the experimental cases 2,15,21 and 35.
- Case C: Information of a the experimental cases 10,17,23 and 38.
- Case D: Mixing of the cases A,B and C.

B. *SVM configuration*

1) Selection of parameter C: In [31], to chose an optimal of regularization parameter C can be derived from standard parameterization of SVM solution given by expressions

$$f(x) = \sum_{i=1}^{n_{vs}} (\alpha_i - \alpha_i^*) K(x_i, x), \tag{11}$$

$$K(x_i, x) = \sum_{j=1}^{m} g_i(x) g_j(x_i), \tag{12}$$

$$\begin{aligned} |f(x)| &\leqslant \left| \sum_{i=1}^{n_{vs}} (\alpha_i - \alpha_i^*) K(x_i, x) \right| \\ &\leqslant \left| \sum_{i=1}^{n_{vs}} (\alpha_i - \alpha_i^*) \right| \cdot |K(x_i, x)| \\ &\leqslant C \cdot |K(x_i, x)| \,. \end{aligned} \tag{13}$$

Further in [31], use kernel functions bounded in the input domain. To simplify presentation, assume RBF kernel function

$$K(x_i, x) = exp\left(-\frac{\left\| x - x^2 \right\|}{2p^2} \right), \tag{14}$$

so that $K(x_i, x) \leqslant 1$. Hence it obtain the following upper bound on SVM regression function:

$$|f(x)| \leqslant \left(C \cdot n_{vs)} \right), \tag{15}$$

Expression (15) is conceptually important, as it relates regularization parameter C and the number of support vectors, for a given value of ε. However, note that the relative number of support vectors depends on the ε -value. In order to estimate the value of C independently of (unknown) s_{vn}, one can robustly let $C \geq f(x)$ for all training samples, which leads to setting C equal to the range of response values of training data. However, such a setting is quite sensitive to the possible presence of outliers, as propose to use it instead the following prescription for regularization parameter:

$$C = max(|\bar{y} + 3\sigma_y|\,, |\bar{y} - 3\sigma_y|), \tag{16}$$

where $\bar{y}$ is the mean of the training responses (outputs), and σ y is the standard deviation of the training response values. Prescription can effectively handle outliers in the training data. In practice, the response values of training data are often scaled so that $\bar{y} = 0$; then the proposed C is $3\sigma_y$.

2) Selection of ε: It is well-known that the value of ε should be proportional to the input noise level, that is $\varepsilon \infty \sigma$. It assume that the standard deviation of noise σ is known or can be estimated from data . However, the choice of ε should also depend on the number of training samples. From standard statistical theory, the variance of observations about the trend line (for linear regression) is:

$$\sigma^2_{y/x} \infty \frac{\sigma^2}{n}. \tag{17}$$

This suggests the following prescription for choosing ε:

$$\varepsilon \infty \frac{\sigma}{\sqrt{n}}. \tag{18}$$

Based on a number of empirical comparisons, it found that works well when the number of samples is small, however for large values of n prescription yields ε -values that are too small. Hence we propose the following (empirical) dependency:

$$\varepsilon = \tau \sigma \sqrt{\frac{ln(n)}{n}}. \tag{19}$$

Based on empirical tuning, the constant value $\tau = 3$ gives good performance for various data set sizes, noise levels and target functions for SVM regression. Thus expression is used in all empirical comparisons as presented in [31].

V. Results

The resulting schemes from the implementation of the proposed discrimination scheme and BSS processes are presented in Figures 6, 7 and 8, where a insight of the behavior of the wavelets coefficients, between the possible combination of emotional could be observed with out the classification process.

Figure 6, presents the characteristics distribution, obtained by the implementation of the proposed strategy by employing the relaxed state (High arousal and low valence(relaxed or calm)) as reference and it comparative to states as anger (Low arousal and high valence (angry or rage)), happiness (High arousal and high valence (happiness or joy)) and sadness (Low arousal and low valence (boredom or sadness).

Figure 7, provides a features distribution representation that implements happiness state as reference to perform a comparative of it behavior with sadness and anger related features and Figure 8, presents the remaining combination of classification process (sadness and anger).

Also the combinations of a greater amount of interlocking elements, to observe the difficulty of the classification problem are also presented in Figure 9.

This representation provides information apriori of the behavior in a features spatial distribution of each emotional state which could be implemented in a classification problem.

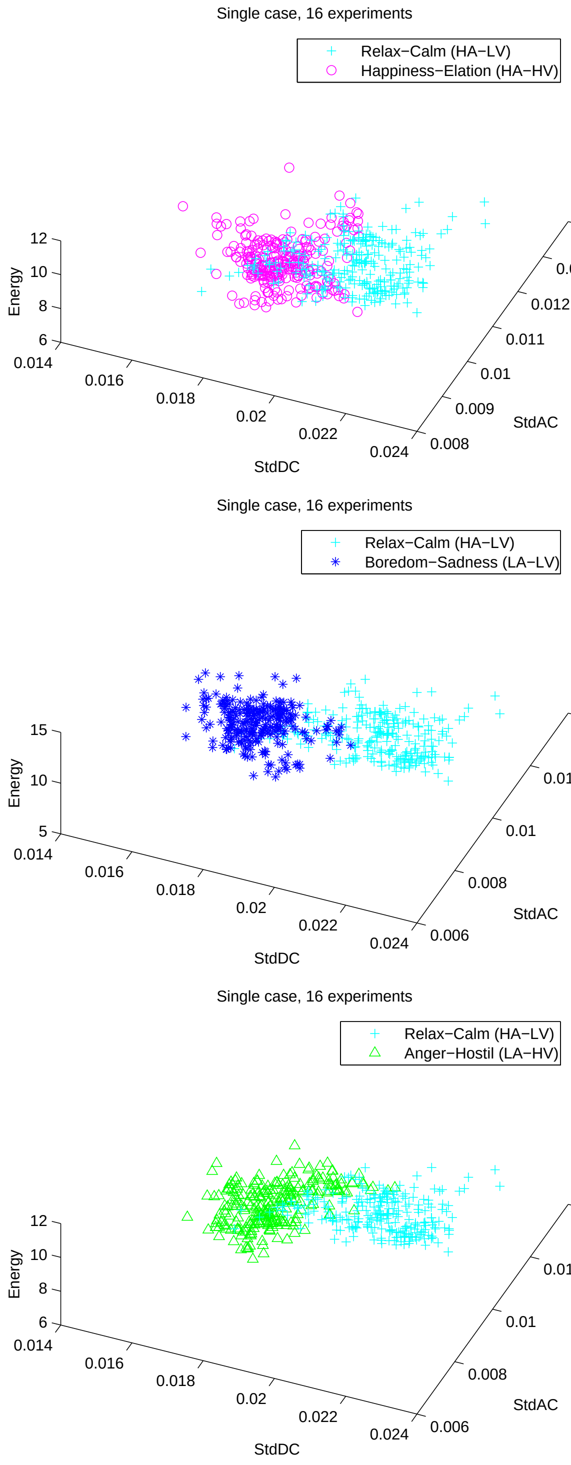


Fig. 6: Features distributed in a Wavelet features distribution, HA/LA as reference.

A. Classification results

The classification stage is performed under the specifications described in Section IV, where each of the emotional states were evaluated independently as shown in the table V.

- Test case A: score an average of 80.75% of classifi-

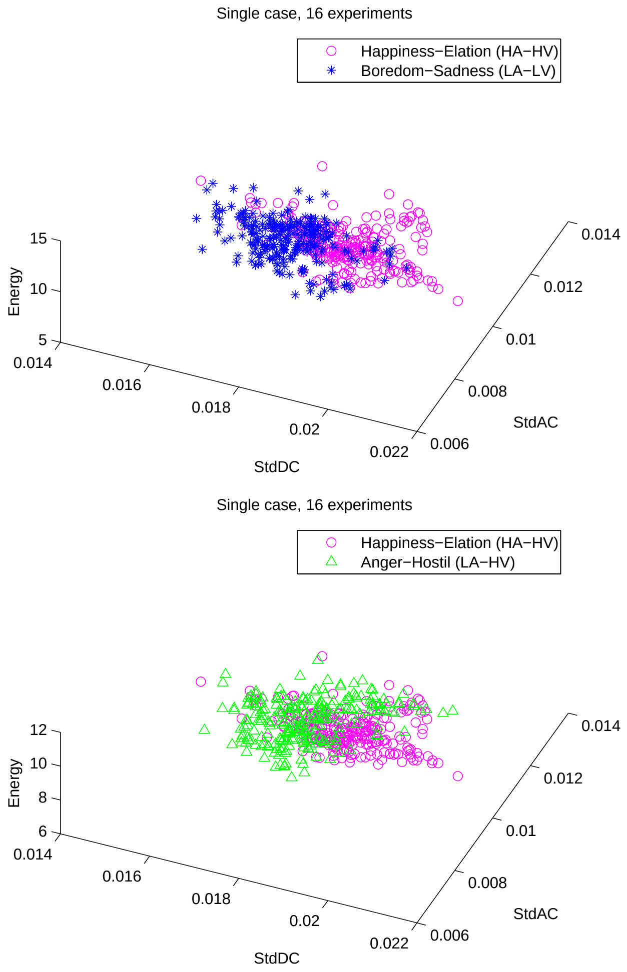


Fig. 7: Wavelet features distribution, HA/HV state as reference

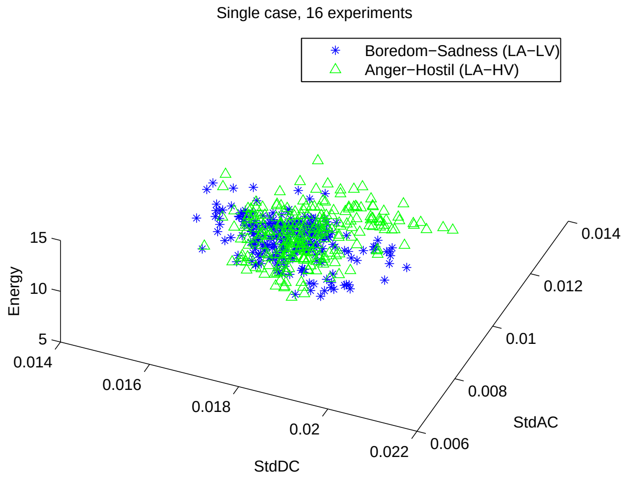


Fig. 8: Wavelet features distribution, LA/LV as reference

cation rate.

- Test case B: score an average of 80.93% of classification rate.

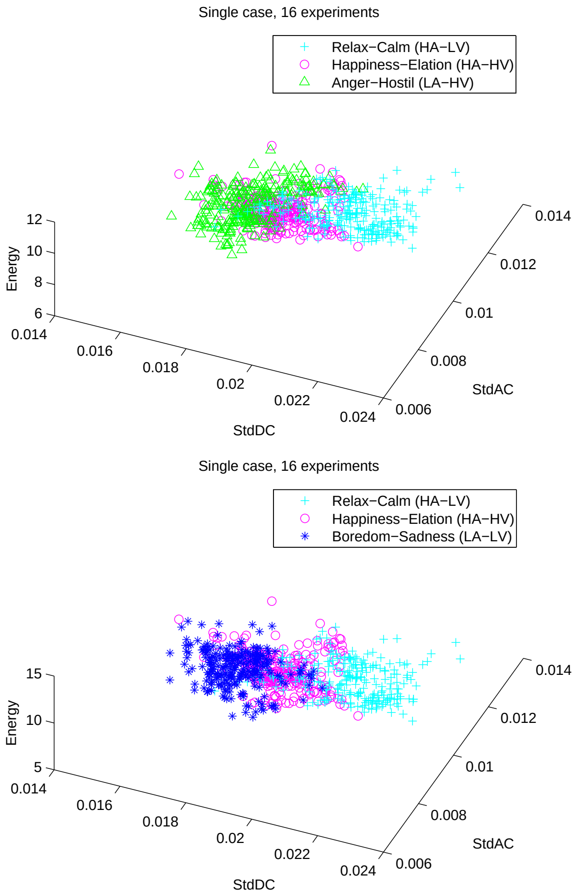


Fig. 9: Features of three emotional states represented as combinations of arousal and valence levels.

- Test case C: score an average of 80.14% of classification rate.
- Test case D: score an average of 86.53% of classification rate.

This results are the average recognition rate of nine hundred experiments and considering only the averages of the cross validation process as show in Figure 10 and obtain a stable recognition rate for the four experimental cases of 82.08% overall performance, in the proposed class variable experimental setups, as shown in figure 11. Furthermore also it shows that with increasing references for the classification process, also increases the percentage in the recognition rate, which implies a correlation between user behavior.

TABLE V: Emotional classification rates by SVM and bounded regions methodology

	Single case	Binary case	Multiclass	
	Classes Recognition rate (%)			
	Single	Two	Three	Four
Case A	77.59	83.61	79.2	82.61
Case B	78.78	83.22	81.19	80.56
Case C	79.61	81.22	78.53	81.23
Case D	84.03	84.85	83.88	83.38

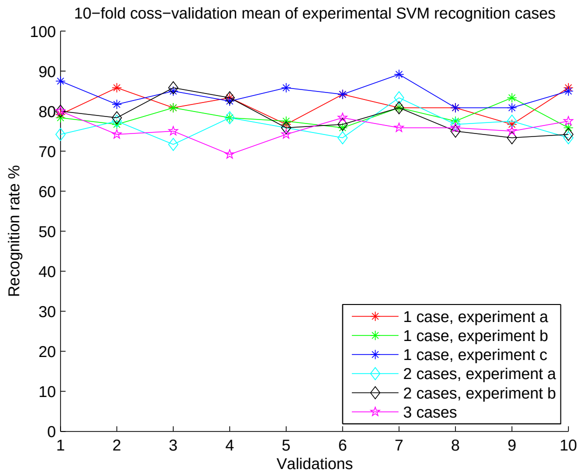


Fig. 10: 10-fold average performance of the svm classification task.

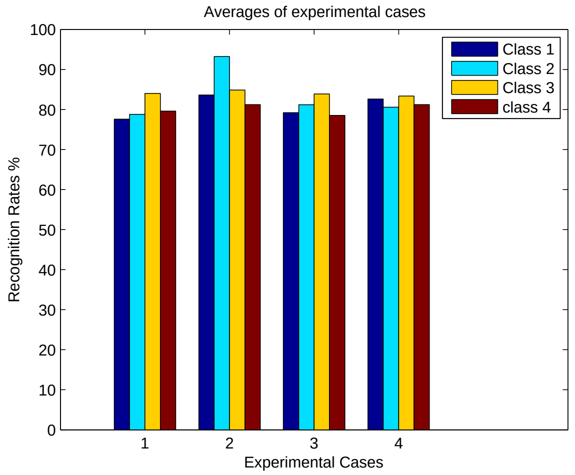


Fig. 11: Overall performance a svm classification task by experimental case.

Then as shown in Figures 6, 7, 8 and 9, if necessary it could develop a classification strategy that considers only valence or arousal states, with a success rate of 90%.

VI. Conclusions

Eventhough the fact that affective computing has shown an important development in recent decades, recognition of emotional states by analyzing biological signals still it presents significant challenges. One of the main problems associated with this type of research is the lack of a standard database and the complexity involved in establishing methodologies of analysis, processing strategy and even an appropriate model of emotional states that are going to investigate.

However, studies such as those presented in this paper, corroborates the existence of a correlation between the electrical activity in the cerebral cortex of people and their moods; and that these differences can be recognized and classified by a process of computational analysis. This paper presents also the feasibility of establishing groups of emotional states to recognize and classify.

On the other hand as the rapid growth and development of sensing techniques and technologies for the analysis of biological signals as well as computational processing strategies facilitate the development techniques for the analysis of mental states. Although wavelet analysis techniques and algorithms classification based support vector machines, other strategies such as search and matching neural networks are used in this paper, they could be employed; as well as a strategy to reduce the data to be processed it could be proposed.

Some of the expectations of implementation for results, is developing a platform for tracking the rehabilitation of persons after an accident or injury, to provide information from the mood of the patient and a specialist can make decisions based on this information .

References

[1] Rosalind W. Picard. *Affective Computing*. MIT Press, 1 edition, 2000.

[2] Saneid Sanei and J.A. Chambers. Eeg signal processing. In *Cardiff University*, 2007.

[3] Yuan-Pin Lin, Chi-Hong Wang, Tien-Lin Wu, Shyh-Kang Jeng, and Jyh-Horng Chen. Multilayer perceptron for eeg signal classification during listening to emotional music. In *TENCON 2007 - 2007 IEEE Region 10 Conference*, pages 1–3, Oct 2007.

[4] P.C. Petrantonakis and L.J. Hadjileontiadis. Eeg-based emotion recognition using hybrid filtering and higher order crossings. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6, Sept 2009.

[5] K. Takahashi. Remarks on svm-based emotion recognition from multimodal bio-potential signals. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pages 95–100, Sept 2004.

[6] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad. Ensemble of svm trees for multimodal emotion recognition. In *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4, Dec 2012.

[7] Y. Attabi and P. Dumouchel. Emotion recognition from speech: Woc-nn and class-interaction. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 126–131, July 2012.

[8] AA Razak, R. Komiya, M. Izani, and Z. Abidin. Comparison between fuzzy and nn method for speech emotion recognition. In *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, volume 1, pages 297–302 vol.1, July 2005.

[9] IJ. Christina and A Milton. Analysis of all pole model to recognize emotions from speech signal. In *Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on*, pages 723–728, March 2012.

[10] M.A Gunler and H. Tora. Emotion classification using hidden layer outputs. In *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on*, pages 1–4, July 2012.

[11] M. Takahashi, O. Kubo, M. Kitamura, and H. Yoshikawa. Neural network for human cognitive state estimation. In *Intelligent Robots and Systems '94. 'Advanced Robotic Systems and the Real World', IROS '94. Proceedings of the IEEE/RSJ/GI International Conference on*, volume 3, pages 2176–2183 vol.3, Sep 1994.

[12] Klaus R. Scherer. What are emotions? and how can they be measured? In *Trends and developments: research on emotions, Social Science Information and SAGE Publications*, volume 44, page 695729, 2005.

[13] Wallace V.; O'Sullivan Maureen; Chan Anthony; Diacoyanni-Tarlatzis Irene; Heider Karl; Krause Rainer; LeCompte William Ayhan; Pitcairn Tom; Ricci-Bitti Pio E.; Scherer Klaus; Tomita Masatoshi; Tzavaras Athanase Ekman, Paul; Friesen. Universals and cultural

differences in the judgments of facial expressions of emotion. In *Journal of Personality and Social Psychology*, volume 4, pages 712–717, 1987.

[14] James A. Russell. A circumplex model of affect. In *Journal of Personality and Social Psychology*, volume 39, pages 1161–1178, 1980.

[15] Bradley M.M. Cuthbert B.N. Lang, P.J. International affective picture system (iaps): Technical manual and affective ratings. In *NIMH Center for the Study of Emotion and Attention, FL: The Center for Research in Psychophysiology, University of Florida*, 1997.

[16] Bradley M.M Cuthbert Lang, P.J. International affective digitized sounds (iads): Stimuli, instruction manual and affective ratings,. In *The Center for Research in Psychophysiology, University of Florida*, 1999.

[17] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M.J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 211–216, April 2006.

[18] G.B. Moody, R.G. Mark, and A.L. Goldberger. Physionet: a web-based resource for the study of physiologic signals. *Engineering in Medicine and Biology Magazine, IEEE*, 20(3):70–75, May 2001.

[19] Stavros Petridis, Brais Martinez, and Maja Pantic. The {MAHNOB} laughter database. *Image and Vision Computing*, 31(2):186 – 202, 2013. Affect Analysis In Continuous Input.

[20] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31, Jan 2012.

[21] V. ; Crystal M. Xiaodan Zhuang, Rozgic. Compact unsupervised eeg response representation for emotion recognition. In *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*, June 2014.

[22] Muhl. C. Soleymani. M. Jong-Seok Lee. Yazdani. A. Ebrahimi T. Pun. T. Nijholt. A. Patras. I. Koelstra, S. Deap: A database for emotion analysis ;using physiological signals,. In *Affective Computing, IEEE Transactions on*, volume 3, pages 18–31, Jan 2012.

[23] S.S.; Chandra S. Arora, S.; Chandel. An efficient multi modal emotion recognition system: Isamc. In *IMpact of E-Technology on US (IMPETUS)*, pages 6–12, Jan 2014.

[24] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, and J.F. Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, April 2013.

[25] Stuart Ira Fox. *Human Physiology*. Mcgraw-Hill (Tx), 7 edition, 2002.

[26] P. Karthikeyan, M. Murugappan, and S. Yaacob. Ecg signals based mental stress assessment using wavelet transform. In *Control System, Computing and Engineering (ICCSCE), 2011 IEEE International Conference on*, pages 258–262, Nov 2011.

[27] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, I Zunaidi, and D. Hazry. Lifting scheme for human emotion recognition using eeg. In *Information Technology, 2008. ITSim 2008. International Symposium on*, volume 2, pages 1–7, Aug 2008.

[28] Lauralee Sherwood. *Human Physiology from cells to system*. Brooks/Cole, Cengage Learning, 8 edition, 2013.

[29] M. Murugappan. Human emotion classification using wavelet transform and knn. In *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on*, volume 1, pages 148–153, June 2011.

[30] S. Dandapat M. Sabarimalai Manikandan. Wavelet energy based diagnostic distortion measure for ecg. In *Biomedical Signal Processing and Control*, volume 2, pages 80–96, Apr 2007.

[31] Yunqian Ma Vladimir Cherkassky. Practical selection of svm parameters and noise estimation for svm regression. In *Neural Networks*, volume 17, pages 113–126, 2004.

Dynamic wireless charging of electric vehicles on the move with Mobile Energy Disseminators

Leandros A. Maglaras
School of Computer Science and Informatics
De Montfort University, Leicester, UK

Jianmin Jiang
Department of Computing
University of Surrey, Guilford, UK

Athanasios Maglaras
Electrical Engineering Department
T.E.I. of Thessaly, Larissa, Greece

Frangiskos V. Topalis
Electrical and Computer Engineering Department
N.T.U.A., Athens, Greece

Sotiris Moschoyiannis
Department of Computing University
of Surrey, Guilford, UK

***Abstract*—Dynamic wireless charging of electric vehicles (EVs) is becoming a preferred method since it enables power exchange between the vehicle and the grid while the vehicle is moving. In this article, we present mobile energy disseminators (MED), a new concept, that can facilitate EVs to extend their range in a typical urban scenario. Our proposed method exploits Inter-Vehicle (IVC) communications in order to eco-route electric vehicles taking advantage of the existence of MEDs. Combining modern communications between vehicles and state of the art technologies on energy transfer, vehicles can extend their travel time without the need for large batteries or extremely costly infrastructure. Furthermore, by applying intelligent decision mechanisms we can further improve the performance of the method.**

***Index Terms*—Electric vehicle; Dynamic Wireless Charging; IVC; Cooperative Mechanisms**

I. Introduction

With regards to the future transport arena, electric vehicles (*EVs*) are considered as the likely replacement of internal combustion engine driven vehicles, especially given the CO_2 reduction and alternative energy perspectives. Electric cars have the potential to reduce carbon emissions, local air pollution and the reliance on imported oil [1]. In Europe, the European commission aims to reduce road transport emissions by *70%* by *2050* [2]. Taking into account the fact that road transport is expected to double by *2050*, passenger cars need to reduce their emissions significantly. Advanced internal combustion engine (*ICE*) technologies are expected to enable emissions reduction, but are not expected to meet long term targets. Electric vehicles, especially plug-in ones (*PEVS*), are penetrating the market and they are currently counted as zero emissions vehicles. Apart from the additional cost of their lithium-ion battery pack that makes them more expensive than conventional vehicles, there are also some other factors that discourage drivers from switching to an *EV*. For instance, electric battery vehicles have a limited driving distance [3] and hence, the current lack of charging infrastructure as well as the total time needed to recharge such a vehicle add to their lack of desirability.

II. Motivation : Increasing a car's all-electric range

In order to surmount this problem, industries and research institutions around the world have proposed numerous solutions. These vary from stationary stations that are scattered across the road network in central positions [4], [5], [6], dynamic wireless charging methods that take advantage of the mobility of nodes [7], [8] and eco-routing algorithms that run in isolation in every vehicle or in a central way for a fleet of vehicles [9], [10], [11]. Dynamic wireless charging is gaining more ground, since it enables power exchange between the vehicle and the grid while the vehicle is moving. Installed infrastructure can be utilized very effectively because many vehicles use the same road segments that are ”equipped” with dynamic charging capabilities. Dynamic charging can take place in a parking lot, at a bus stop during passenger disembarkation, along a highway or near traffic lights.

A. Wired charging

Electric vehicles are plugged for charging on the existing electrical grid infrastructure, but sometimes the electrical infrastructure is inadequate for supporting this additional energy demand of high power fast charging stations. Moreover, the presence of several concurrent charging requests could cause overload conditions in local nodes of the grid, if the charging processes of the PEVs are not properly managed and scheduled. One alternative to fast charging stations [4] is to have mobile charging systems (*MCSs*)with a high storage Capacity and a mobile charging system for electric vehicles is presented in [12]. These stations can be a solution when the electrical infrastructure of the local grid is unable to support high power fast charging stations.

Smart scheduling strategies can be profitably used to manage the (*PEV*) charging problem [5], for based on quadratic programming for charging *PEVs*, these can decrease the peak load and flatten the overall load profile. The usage of Information and Communication Technology (*ICT*) in a smart grid environment is a proposed solution [13], [14]. Regarding which, the authors in [15] advocate the deployment of smart grid communication architectures by using small embedded systems in a hierarchical way or a manner that can enable the distribution grid to charge a large number of *EVs* without the need to carry a high workload.

B. Dynamic wireless charging

Dynamic wireless charging is gaining more ground, since it enables power exchange between the vehicle and the grid while the vehicle is moving.

Recently, the Telewatt project introduced an original approach involving the reusing of existing public lighting infrastructures for such charging, whereby a fraction of the power not consumed by lamps at night can be used for the benefit of the charging stations. The service is accessible by a smartphone application, where clients specify to the TeleWatt server their destination and their battery level, for which they receive a response of a list of available charging terminals close to this destination [6]. Hevo announced a novel dynamic charging system where manhole covers will be used as charging stations and a pilot program is scheduled to be performed in New York in 2014. Two Online Electric Vehicle (*OLEV*) buses that can charge during travel have been put into service for the first time in the world on normal roads in the city of Gumi - Korea by the Korea Advanced Institute of Science and Technology (KAIST) [7]. The power is transmitted through magnetic fields embedded in the roads. That is, it comes from the electrical cables buried under the surface of the road, thus creating these magnetic fields and the length of power strips installed is generally 5%-15% of the entire road. In [16] the authors present a method for power transfer between electric vehicles, where drivers ”share” charge with each other using the inductive power transfer (*IPT*) of the charge between vehicles at rendezvous points. However, one major issue with this concept is the technology requirements that have to be met by passenger vehicles in order for this solution to be feasible. Moreover, the dynamic charging of vehicles raises health issues related to the leaking magnetic flux from IPT.

C. The charging station location problem

Thoughtful siting of public charging stations can ease consumer range anxiety while offering a lower cost approach to integrating *EVs* into the transportation market. The authors in [17] propose a method to anticipate parking demands and more efficiently to locate *EV* charging infrastructure in new settings and/or subject to different constraints. The researchers in [18] used Lisbon, Portugal, as a case study where they determined not just the locations to be installed, but also their capacity of at each location, with the aim of optimizing the demand covered within an acceptable level of service. In [11], the authors try to build a comprehensive objective function taking into account geographic information, construction cost and running cost in order to achieve optimal planning of charging stations. In [19] an electric vehicle battery swapping station is described as well as a business case scenario provided where customers have access to the stations such that they can meet their motion energy requirements by swapping batteries for charged ones quite quickly.

D. Eco-routing of EVs

Similar to eco-routing for conventional vehicles, novel methods are being developed and used in order to reduce the energy consumption of *EVs* [9], [20]. The authors in [9] have developed an eco-routing navigation system, which determines the most eco-friendly route between a trip’s origin and its destination. With the use of a Dynamic Roadway Network database that integrates historical and real time data they managed to reduce CO_2 emissions. Based on their previous work they now aim to create an eco-routing algorithm that will be incorporated into a prototype eco-routing navigation system for*EVs*. In [20], they moved a step further by creating a routing system that could extend the driving range of *EVs* by calculating the minimum energy route to a destination.

E. Contributions

The present article develops a cooperative mechanism for dynamic wireless charging of electric vehicles and makes the following contributions:

- The concept of mobile energy disseminators is introduced
- A cooperative mechanism based on inter-vehicular communications (*IVC*) and long term evolution (*LTE*) communication is proposed
- A route optimization problem for every electric vehicle is formulated
- The proposed mechanism is evaluated through extensive simulations

The rest of this paper is organized as follows: Section III describes the concept of mobile energy disseminators (MED) and in Section IV the communication mechanisms are presented. In Section V the optimization problem is formulated, whilst Section VI presents the simulation environment And the results. Section VII concludes the article.

III. Mobile energy disseminators

Energy exchange can be facilitated by inductive power transfer (*IPT*) between vehicles and/or by installing a roadside infrastructure unit for wireless charging. However, given the vast expanse of road networks, it is impractical to have infrastructure units on every road segment due to prohibitive costs. *IPT* allows efficient and real-time energy exchange where vehicles can play an active role in the energy exchange procedure. On the other hand, the use of mobile nodes as relay nodes is common in vehicular ad hoc networks (VANETs). In a VANET mobile relay nodes can serve as carriers and disseminators of useful information. Influential spreaders, nodes that can disseminate the information to a large part of the network effectively, are an open issue in ad hoc networks. That is, nodes with predefined or repeating routes that can cover a wide range of a city region can do the work of roadside units while exploiting their mobility in order to provide higher quality-of-service (*QoS*).

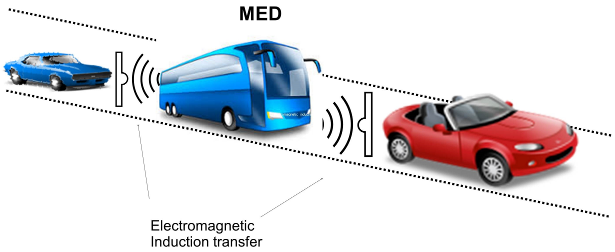


Fig. 1. Electromagnetic induction transfer

Similar to information dissemination, special nodes, like buses (trucks), can act as energy sources to *EVs* that need charging, in order to increase travel time. These vehicles, form now on called mobile energy disseminators (*MEDs*), use electric plug in connection or IPT in order to refill starving EVs. Buses can play the role of *MEDs* since they follow predefined scheduled routes and their paths cover a major part of a city, while trucks could have the role of energy chargers mainly on highways. Buses can be fully charged when parked, before beginning their scheduled trip, and can be continuously charged along their journey by *IPT* stations installed at bus stops. Additional technology requirements that these vehicles may need in order to operate as energy sources, is an open issue, but it is rather more feasible in the near future, to have these features installed into large public vehicles than into passenger vehicles due to the additional cost and space requirements. Vehicles that book charging places on the same *MED* can create clusters and mobile charging stations will play the role of the clusterheads.

The vehicle requiring electric charge will approach the appropriate truck, after a preceding agreement, from the rear or the front end depending on the vehicle construction. The procedure will provide vehicle charging by an electric plug in connection (or process), or by electromagnetic induction with the use of Tesla coils. Immobilized charging can take place at predetermined road points (for example parking areas) in order to avoid traffic obstruction and in this case the method of the plug in electric connection is preferable. A synchronization of the vehicles' movement will be executed via wireless communication mainly controlled by the truck/bus. From the analysis undertaken, it is apparent that it is preferable, for reasons of safety and better management of the system, that the vehicle needing charge should move ahead or behind the truck creating a cluster [21], [22] or platoon [23]. There will be a special joint magnetic arrangement concerning the vehicles, as well as a special interlocking arrangement in order for the two vehicles to approach and remain in contact, even while in motion, for as long as the charge transfer takes place. Charge transfer can be achieved with electric plug in connection, or by electromagnetic induction. During the latter transfer, the charge and consequently the power transfer will be accomplished with the use of two detached subsystems of magnetic coupling of high efficiency.

The electromagnetic subsystems will include magnetic coils, which will cooperate and function like the primary and the secondary coils of a transformer, which will have loose coupling using air as the proper medium. This way of coupling (like Tesla coils) has proven to be more efficient than using ferromagnetic materials. The primary coil of the truck will be movable and able to be inserted in the bigger diameter coil of the vehicle, in order to improve the efficiency factor of the power transfer process and to minimize the leaking of magnetic flux. Moreover, the two subsystems will be specially shielded (Faraday cage) in order to protect occupants and bystander vehicles or pedestrians from electromagnetic radiation. The truck/bus will carry high capacity batteries and if needed, the appropriate electric system to convert voltage from *DC* to *AC* voltage of high frequency. It will also need to carry a conventional internal combustion engine in addition to the correct electric generator, to be used to produce electric energy in an emergency situation.

The advantages of the proposed system are a) high efficiency factor (especially when the charge transfer is achieved via an electrical plug in connection) b) very short delay regarding the moving of the vehicles c) significant reduction of environmental pollution and d) coverage of special needs in exceptional climatic conditions or failure conditions.

IV. Cooperative mechanisms

By using cooperative mechanisms, based on dedicated short-range communication (*DSRC*) capabilities of vehicles or long term evolution (*LTE*) technology, vehicles search for the *MEDs* in range and arrange a charging appointment while moving [8]. In the following section, we present a cooperative mechanism, based on the($DSRC$) and LTE capabilities of vehicles. A network $G = (N, L)$, where N is the set of nodes (intersections) and L is the set of links (road segments), is considered. V is the set of electric vehicles that move in the network and M is the subset of electric vehicles that can act as mobile charging stations.

A. IVC system

In order to state its presence, each MED periodically broadcasts cooperative awareness messages (*CAM*). Each beacon message consists of a node identifier (V_{id}), node location, scheduled trip (a subset of set L), current charging capability (CC) and energy value (E/KWh). CC is the current energy that the mobile charging station can afford to dispose of to charge the vehicle without jeopardizing its own needs. These messages are disseminated by all vehicles that act as relay nodes.

Each EV that needs energy, upon receiving a CAM by an MED performs the following steps:

1) Checks whether MED is on his route or not according to their current positions and destinations
2) Checks whether the CC level is high enough in order to cover its energy needs
3) Asks for a charging place by sending a CAM which contains minimum charging time

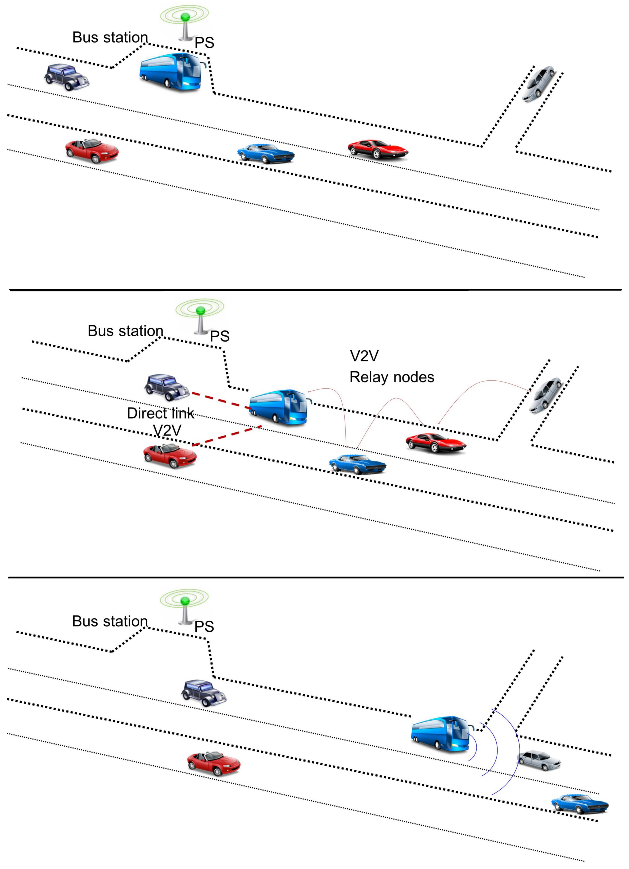


Fig. 2. Application example of a mobile energy disseminator: A. Contactless charging is used to deliver charging to a bus B. V2V communication between MED and EV C. EV recharges from the bus using IPT

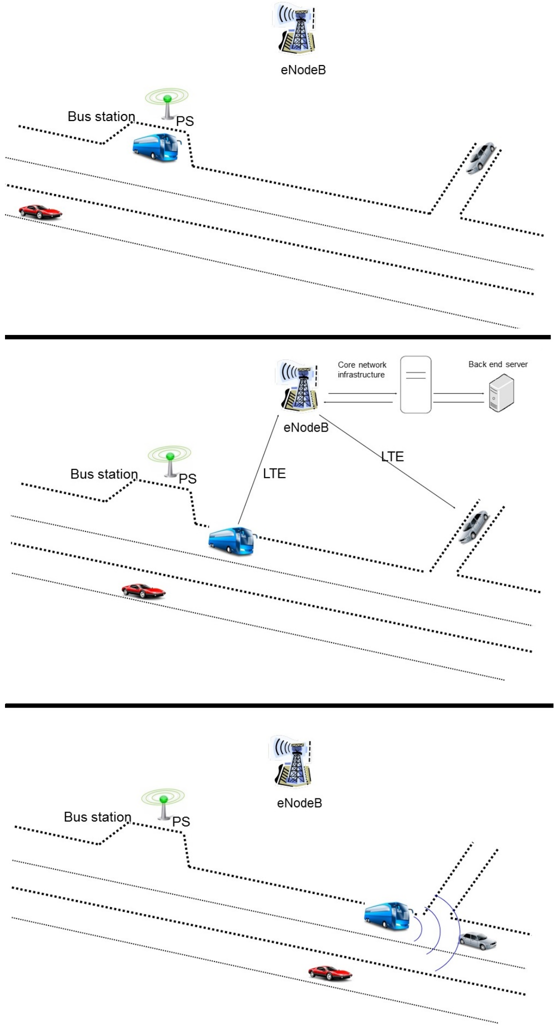


Fig. 3. Application example of Mobile Energy Disseminator: A. Contactless charging is used to deliver charging to a bus B. LTE communication between MED and EV C. Electric vehicle recharges from the bus with inductive power transfer (IPT)

4) Chooses to select this bus as the wireless energy transfer station
5) Books a charging place
6) Drives in front of or behind the bus for the determined time period in order to recharge

Steps 3-5 constitute the negotiation phase, in which MED and EV exchange dedicated short range messages $(DSRC)$ in order to confirm the energy transfer. An assumption that we make is that vehicles can book their charge of battery as soon as they realize that their charge level is low and a MED meets their criteria on relative distance and available energy. The architecture of the proposed mobile energy dissemination is demonstrated in Figure 2.

B. LTE system

For the LTE system, we assume that vehicles are equipped with The Evolved Universal Terrestrial Radio Access Network $(EUTRAN)$ interface, which enables the vehicles to communicate with the eNB so as to access the core components of the LTE. LTE Evolved Node B (eNB) base station transceivers are deployed alongside the road network in order to cover the area. Each bus communicates to the LTE the scheduled trip that is going to be followed, the available charging capability and energy value and charging availability, similar to the IVC system. All vehicles are assumed to be equipped with GPS.

Each EV that needs energy:

1) Checks whether MED is on his route or not according to their current positions and destinations
2) Checks whether the CC level is high enough in order to cover its energy needs
3) Checks whether the MED is already fully booked
4) Books a charging place
5) Drives along the bus for the determined time in order to recharge

The architecture of the proposed mobile energy dissemination architecture is demonstrated in Figure 3. The benefits of such an approach are threefold: First, it utilizes existing cellular infrastructure. Second, the $802.11pnetwork$ overhead introduced by frequent communication between EVs and

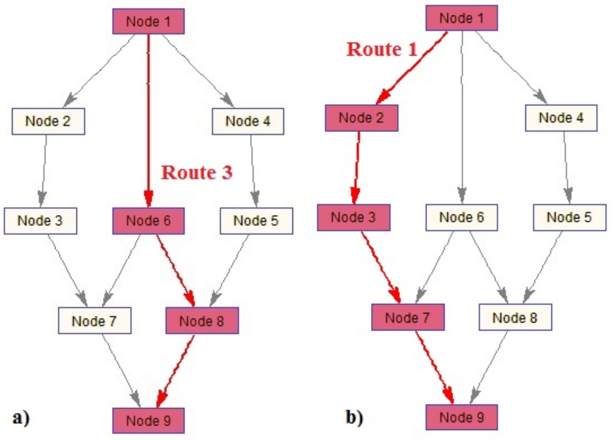


Fig. 4. Optimum path selection based on a) Distance (Route 3), b) Time (Route 1)

$MEDs$ is offloaded. Third, information is more up to date than that received through IVC, where many intermediate relay nodes may be needed in order to disseminate data effectively. However, vehicles are required to have two types of network interface cards. Moreover packets that pass through the LTE core potentially experience more delay.

Route selection algorithms, where vehicles communicate with each other in order to exchange information are crucial in order to evaluate the performance of the method. Optimal route selection overcomes a common problem in which all vehicles are preferring the same paths, leading to over congestion. Optimal routing of vehicles that use this new technology can be formulated as a modified shortest-path problem where the weights of the road segments may vary over time, according to the existence or not of a *MED* traveling on the road segment [24].

V. Formulation of the Method

Vehicle routing aims at identifying a feasible route that satisfies metrics constraints. These metrics are associated with multiple parameters which can include delay, distance and energy cost. Mathematical formulation of vehicle routing is intrinsically the restricted shortest path problem (RSP). In order to test the performance of the method we solve an optimization problem where an objective function that combines time, distance and power is used in order to make the best route selection. The optimization problem is solved in isolation for each vehicle based on whether the road segments include $MEDs$, the additional distance that the vehicles have to travel and the predicted travel time to the destination. A simple graph is shown in Figure 4, where a vehicle has to choose between four routes, one containing a MED and three one of which being the shortest path in terms of distance

The optimization problem for each vehicle i is formulated:

$$min \sum_{jk} W_{ijk}, \; where \; W_{ijk} = F(E_{ijk}, T_{ijk}, D_{ijk}) \quad (1)$$

s.t.

$$\sum_{jk} T_{ijk} \leq T_{th} \quad (2)$$

$$\sum_{jk} D_{ijk} \leq D_{th} \quad (3)$$

$$\sum_{jk} E_{ijk} \leq E_o + E_{ind} \quad (4)$$

where, W_{ijk} is the weight assigned to each road segment jk of the route of vehicle i. E_{ij} is the energy that is being consumed on road segment jk, E_o is the initial energy of vehicle and E_{ind} is the induced energy to the vehicle i. Constraints *(2)* and *(3)* are used in order to avoid sending all vehicles from routes that are too long, which would lead to excessive cost in terms of time and distance of the trip. T_{th},D_{th} are parameters that give the upper limit of the overall time and the distance that a vehicle can spend in order to reach its destination. The energy cost of every road segment can be expressed as a proportion of the mean velocity. The velocity is the quotient of the distance of the road segment and the time that the vehicle will need to spend on this segment, on average. The two forces that oppose the motion of an automobile are rolling friction, F_{roll} and air resistance,F_{air}.

$$F_{roll} = \mu_{\tau} * m * g, \;\; F_{air} = \frac{1}{2} A * C * p * u^2 \quad (5)$$

where, m is the mass of the car in Kg, $g = 9.8 m/s^2$, u is the mean velocity in m/s and μ_{τ} is the rolling resistance coefficient. C is a dimensionless constant called the drag coefficient that depends on the shape of the moving body, A is the silhouette area of the car (m^2) and p is the density of the air (about 1.2 kg/m^3 at sea level at ordinary temperatures). Typical values of C for cars range from *0.35* to *0.50*.

In constant-speed driving on a level road, the sum of F_{roll} and F_{air} must be just balanced by the forward force supplied by the drive wheels. The power that a vehicle needs when traveling with a steady speed is given by Equation 6.

$$P = \eta * F_{Forward} * u = \eta(F_{roll} + F_{air}) * u \quad (6)$$

,η is the efficiency factor of the system.

The energy cost of vehicle i for traveling in road segment (j), E_{ij}, is calculated by Equation 7.

$$E_{ij} = P * T_{ijk} \quad (7)$$

If the road segment belongs to the path of a MED, then the vehicle can increase its energy by induction. The amount of the induced energy is proportional to the total time that the EV and the MED will stay connected. This time depends on the meeting point between the vehicle and the MED in relation to the total road segment length and the availability of the MED. In order to represent the induced energy to the electric vehicle Equation 7 is rewritten:

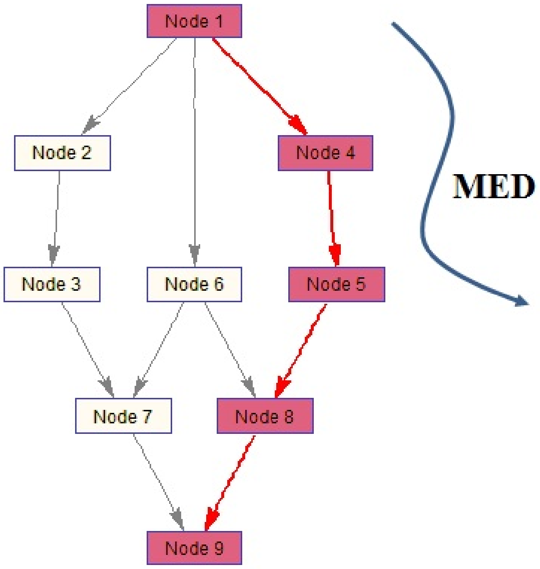


Fig. 5. Optimum path selection based on energy (*Route 4*)

$$E_{ijk} = P * T_{ijk} - E_{ind} \tag{8}$$

where, E_{ind} is given by:

$$E_{ind} = t_{cont} * C_{ind} * P_{ind} \tag{9}$$

C_{ind} is the induction coefficient and t_{cont} the time of contact between the MED and the EV. P_{ind} is the power of the MED. We ignore acceleration and deceleration phenomena. The edge parameters for the example network of Figure 4 are shown in Table I. The mean energy induced from a MED to a vehicle is calculated according to Equation 9 and the hypothesis that all vehicles that traverse the right part of the graph can be in contact with the MED throughout the time that it takes in order to traverse each road segment. Time of contact between the MED and EV is calculated according to the length of the road segment D_{jk} and the velocity of the MED, u_{med} (See Equation 10).

$$t_{cond} = \sum_{jk \in P'} \frac{D_{jk}}{u_{med}} \tag{10}$$

Where, P' is the path that MED and EV share, and it is constituted by individual road segments.

For the simulated scenario presented in Figures 4 and 5, Table I presents edge parameters. By comparing columns of time, distance and energy cost, it is obvious that the overall cost of each road segment is different according to the evaluation metric used. In the table, the optimum route for every single parameter (time, distance and energy) is highlighted and it is obvious that for the trip from node *1* to node *9* the optimal route in terms of energy consumption, without the existence of a MED, is Route *3* (Nodes: *1, 6, 8, 9*). In the presence of a MED that travels road segments (*1*,*4*), (*45*) and part of road segment (*58*) the energy cost of Route *4* (Nodes: *1, 4, 5, 8, 9*) decreases, thus making it the optimal route

TABLE I
CALCULATION OF DISTANCE, TIME AND ENERGY COST FOR EVERY ROAD SEGMENT AND ROUTE

Road ID	*Dist (Km)*	*Time (min)*	*Vel (Km/h)*	*E cost (kWh)*	*E ind (kWh)*	*Total E (kWh)*
12	8	17	28.24	1.315	0.00	1.315
24	8	24	20.00	1.315	1.200	0.114
16	20	67	17.91	3.286	0.000	3.286
23	12	17	42.35	1.972	0.000	1.972
37	12	17	42.35	1.972	0.000	1.972
45	12	33	21.82	1.972	1.800	0.172
58	16	33	29.09	2.629	1.200	1.429
67	12	25	28.00	1.972	0.000	1.972
68	8	30	16.00	1.314	0.000	1.314
79	4	17	14.12	0.657	0.000	0.657
89	6	33	10.91	0.986	0.000	0.986
Route						
1	*36*	**68**		*5.920*		*5.920*
2	*36*	*109*		*5.920*		*5.920*
3	**34**	*130*		*5.587*		*5.587*
4	*42*	*123*		*6.920*	4.200	**2.702**

The objective function can be any combination of the parameters time, distance and energy. Using a linear combination with different weights (w_i) per parameter (see Equation 11) is a solution that produces satisfactory outcomes and can be tuned in order to favour time, distance, energy or any combination of the parameters, according to the preferences of the driver.

$$W_{ijk} = w_1 * |E_{ijk}| + w_2 * |T_{ijk}| + w_3 * |Dijk| \tag{11}$$

The road segments that comprise the routes that each vehicle can follow are:

- Route 1: Nodes (1, 2, 3, 7, 9)
- Route 2: Nodes (1, 6, 7, 9)
- Route 3: Nodes (1, 6, 8, 9)
- Route 4: Nodes (1, 4, 5, 8, 9)

VI. METHOD EVALUATION

In order to evaluate the effect of the routing method to the EV total range we conducted simulations where vehicles are injected into the road network. Vehicles choose their path according to distance, time and energy cost. In addition, the vehicles entering the system have initial energy according to a uniform distribution and the energy of each is measured when it reaches *node 9* (exit node). All nodes are equipped with GPS receivers and On Board Units (OBU) and the location information of all vehicles/nodes, needed for the clustering algorithm, is collected with the help of the receivers. The only communications paths available are via the ad-hoc network and there is no other communication infrastructure. The power of the antenna is P_{tx} = 18dBm and the communication frequency f is *5.9* Ghz. In our simulations, we use a minimum sensitivity (P_{th}) of -69 dBm to -85 db, which gives a transmission range of *130* to *300* meters.

TABLE II
MINIMUM SENSITIVITY IN RECEIVER ANTENNA ACCORDING TO DATA RATE.

Data Rate (Mb/sec)	*Minimum Sensitivity(dBm)*
3	-85
4.5	-84
6	-82
9	-80
12	-77
18	-70
24	-69
27	-67

TABLE III
SIMULATION PARAMETERS

Independent parameter	*Range of values*	*Default value*
$Number\ of\ MEDs$	1-2	1
$Number\ of\ vehicles$	50-100	100
$Initialenergy$	4-24 kW	24 kW
$MED\ capacity$	1-2	2
C_{ind}	0.7-0.8	0.8
P_{ind}	20-50 kW	40 kW
η	0.7-0.8	0.8
E_{th}	6 -10 kW	6 kW
$Vehicle\ rate(eh/min)$	1/5 - 1/15	1/5
$Parameters\ w_1, w_2, w_3$	0-1	1/3

The arrival rate of the vehicles follows the Poison process with parameter λ and the speed assigned to them is associated with the speed limit of each road segment. The range of values is given in Table III. Each MED moves in circles representing the path that a bus follows during a day in the city and each vehicle in need of energy sends its query using a broadcast mode. Each MED that receives the query replies with its availability according to the procedure described in subsection IV-A. In the case where a vehicle receives more than one reply from different $MEDs$, it decides which to choose according to the objective function 11 and the values of the weights w_i. Each MED stores the received applications and final decisions from the EVs along with the unique ID that each vehicle has.

In the upcoming subsections, each figure represent a snapshot of a simulated scenario except for figures 11, 17, 18 and 19 as well as tables IV and VI, where the mean values of the outcomes of 100 different simulations are represented.

A. Network with one MED

In the first evaluation scenario, we simulate the network of Figure 5, where only one MED exists. Vehicles decide to follow the energy efficient, time efficient or the distance efficient path (shortest path), according to the policy of the driver. Based on this policy, the weights w_1, w_2, w_3 assigned to time, distance and energy are different and the vehicles follow different paths.

The induced energy to the vehicles that traverse road segments (*12, 45, and 58*) is given by Equation 9. It is worth mentioning that in all the simulated scenarios the vehicles wait at the beginning of the road segment until the MED approaches the intersection so as to achieve maximum contact time. Therefore, the vehicles choose to wait for the MED to arrive even if it is not near and once in contact accept a longer path as necessary. This behaviour may lead to some vehicles having additional energy consumption, due to longer trip in terms of time. Moreover, this situation, where vehicles wait in the beginning of the area when a MED is moving, induces additional parameters in the routing algorithm, e.g. waiting cost that is translated into additional energy cost, time cost and additional total load in the network due to stationary cars. In the simulation environment $MEDs$ have the capability of charging up to two vehicles simultaneously and they ignore any incoming calls for charging when they are fully booked. The simulation parameters are given in Table III.

1) Energy efficient scenario: In this case, each vehicle V_i that enters the simulation area, decides upon the optimal route based on an energy threshold, which can be derived from Equation 4. If initial energy is below *6kW*, which is the maximum amount needed in order to reach the destination (*node 9*), the vehicle asks the MED for an empty induction slot, which has two places of charging. If it has an empty slot then V_i books it and waits for the MED to reach node *1* in order to begin recharging (*Route 4*). In the case where the initial energy of the vehicle is above threshold or no free charging place exists, it chooses to follow the shortest path in terms of distance (*Route 3*). Each MED is driving on a ring road which consists of road segments *(14), (45)*, part of road segment *(58)* and other roads out of our simulation area of the same distance. The MED is driving at a velocity that is *90%* of the maximum for the road segment, whilst the EVs travel at *100%* of the allowed speed limit. When a vehicle V_i follows a MED for recharging its velocity drops to the latter's velocity.

In Figure 6, we observe the energy consumption of each vehicle during its travel in the simulation area. Vehicles enter the simulation area with a random initial energy between *4* and *24kW*. Vehicles with low initial energy, book a free charging place on the *MED* and manage to lower the consumed energy in the area.

Vehicles that choose to recharge during their travel in the simulation area need to wait for the MED to reach node *1*. Bearing in mind that ***MED*** is following a different route, which includes road segments outside of the simulated area, and the current position of the MED at the time that the booking application happens, the vehicle calculates the waiting time and decides whether to wait or not for the MED. In this scenario vehicles do not have any time threshold, and wait no matter how long the time may be, as long as the MED has a free charging place. This additional waiting time is represented in Figure 7.

When a vehicle V_i follows the MED in order to recharge, in addition to more time this choice leads to a greater distance having to be traveled. This additional distance is represented in Figure 8.

2) Impact of the time threshold: In this scenario, vehicles with energy below the threshold (*6 kW*) decide to recharge only

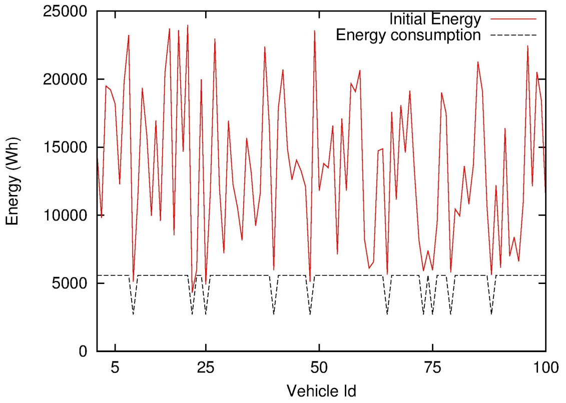


Fig. 6. Energy consumption of Vehicles

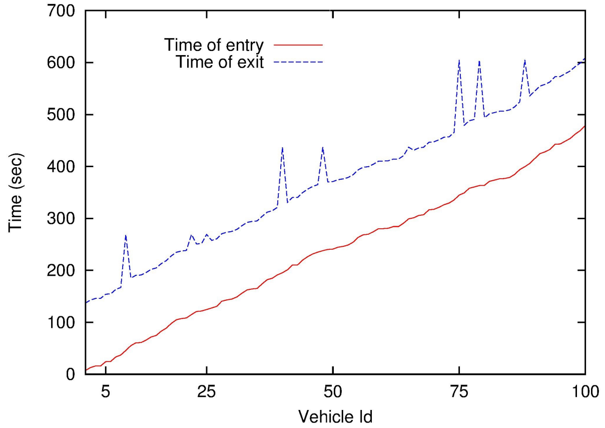


Fig. 7. Travel time of vehicles

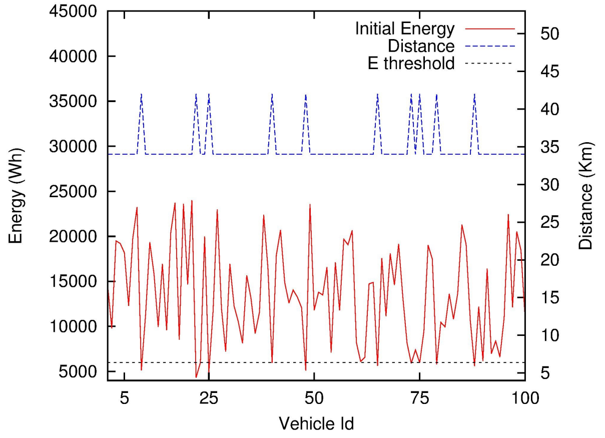


Fig. 8. Distance covered by every vehicle in the simulation scenario

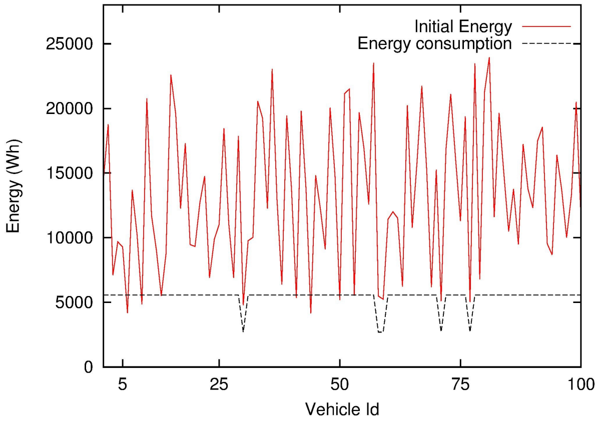


Fig. 9. Energy consumption of vehicles with time threshold

if this procedure is not going to induce too much additional travel time and in order to achieve this, every has its own maximum wait time for the MED. When recharge is needed, a vehicle calculates the required wait time and if total travel time does not exceed the time threshold then it chooses to recharge (*Route 4*). Conversely, if the total travel time exceeds the time threshold, the vehicle decides to follow the shortest path in terms of distance (*Route 3*) and the effects of this policy on energy are depicted in Figure 9. In the simulations conducted the time threshold is assigned a value that is the same as the time that a MED needs in order to perform a circle. It is assured that each vehicle waits for the MED to complete at most one circle until it reaches the initial point of charge (*Node 1*). In the case when the approaching MED has no empty charging spot, the vehicles do not wait for another circle to be completed and follow the shortest path and it is assumed that they find a stationary charging spot along their way in order to refill.

3) Impact of E threshold: Increasing the energy threshold causes an increase in the demands for recharging from starving vehicles. Since MED has a limited capacity (in this experiment the capacity of the MED has the default value *2*), large energy thresholds will not have any positive effects on the overall performance of the system. Vehicles will either need to wait for longer times or will not be able to find a free charging place (Figures 10, 11). If a MED has an empty slot, then V_i books it and waits for the former to reach node *1* in order to begin recharging (*Route 4*). In the case that the initial energy of vehicle is above threshold or no free charging place exists, it chooses to follow the shortest path in terms of distance (*Route 3*).

We can observe in 10, that vehicle *25* enters the area with low energy, but since no free charging place is available on the MED it runs out of power and the same happens with vehicles *65, 73 and 88*. Increasing the threshold but keeping the number of charging places to the default value of *2*, which is the feasible scenario, the number of vehicles that can be served does not increase. At the same time, vehicles with no

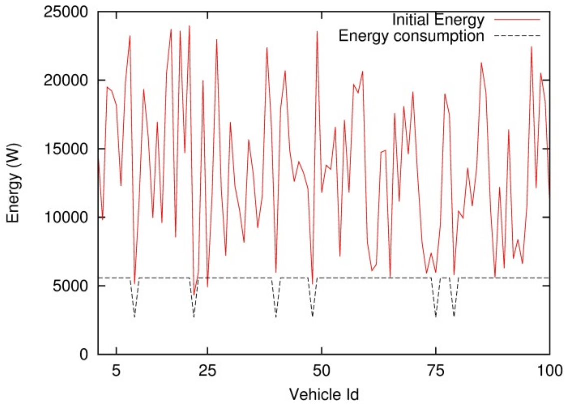


Fig. 10. Energy consumption of vehicles. MED with *2* charging places, Energy threshold *6kW*

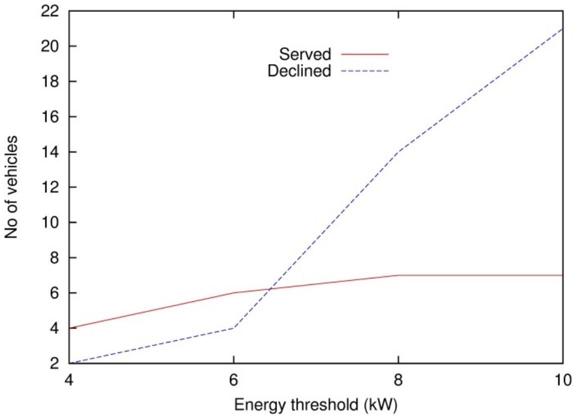


Fig. 11. An increase in the energy threshold leads to more requests for charging and since $MEDs$ capacity is limited most requests are declined.

immediate need for energy may occupy a charging place, thus leaving out vehicles that are starving (Figure 11).

4) Impact on the vehicles' total range: The initial purpose of the method is to increase the mean total range of electric vehicles. Based on the above scenarios, we can deduce that since the energy of starving vehicles is protected the mean total range of every vehicle of the investigated system will also increase. In order to visualize this effect, we extract the remaining energy of all the vehicles after they reach node *9* and compare it with that rmaining if they all followed the shortest or the fastest path.

In Figure 12, we present the remaining energy of every vehicle after reaching node *9* and it can be seen that only one runs out of energy before reaching its destination, when using the energy optimal efficient scenario. That is, this vehicle does not manage to find an empty place for charging on the MED and the initial energy that it had was not sufficient for it to reach its destination. In the shortest path and fastest time scenarios the number of vehicles that run out of energy and have to stop before reaching the final destination is given in Table IV. The same table also shows the mean value of the remaining energy. It is evident that the energy efficient scenario, based on the use of $MEDs$, gives better outcomes and balances the consumed energy of the vehicles.

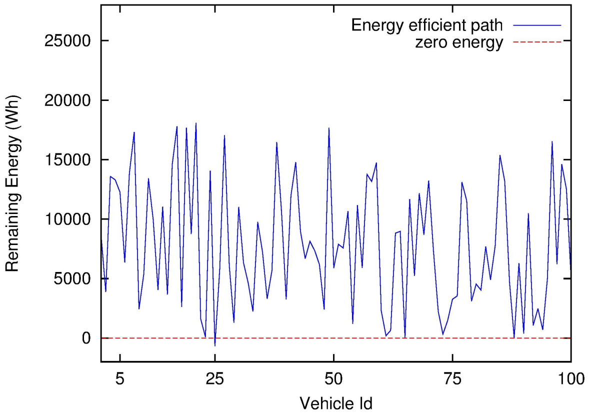


Fig. 12. Remaining energy of vehicles when reaching destination node *9*

TABLE IV
EVALUATION OF PERFORMANCE BASED ON REMAINING ENERGY E OF VEHICLES AFTER HAVING REACHED TARGET NODE *9*.

Method	*No of Vehicles with with* $E_{out} < 0$	*Mean* E_{out} kWh	*Additional range %*
$Energy\ path$	1	10.096	
$Shortest\ path$	4	9.072	-11.3 %
$Fastest\ path$	9	7.044	-43.3 %

Based on the remaining energy, we can easily find the total additional range that each vehicle can travel and in fact, $MEDs$ manage to increase the total range of each vehicle in a typical urban scenario by as much as 43.3

B. Network with two MEDS

In the second evaluation scenario the network we use is the one from Figure 13. The graph includes a second MED that induces further added distance to a vehicle that chooses to follow it. Vehicles decide to follow the energy efficient, the time efficient or the distance efficient path (shortest path) according to the policy of the driver.

The new added route *5* comprises the nodes *(1,2,10,11,7,9)*.

1) Energy efficient path: If the initial energy of a vehicle is below *6kW*, which is the maximum amount of energy needed in order to reach its destination, it asks both $MEDs$ for an empty energy induction slot. Each MED has two places for charging. If one has an empty slot, then V_i books this slot and waits for the MED to reach the closest node in order to begin recharging (*Route 4 or Route 5*). In case that the initial energy of vehicle is above threshold or no free charging place exists, vehicle chooses to follow the shortest path in terms of distance (*Route 3*).

In Figure 14, we observe the energy consumption of each vehicle during its travel in the simulation area. Vehicles enter

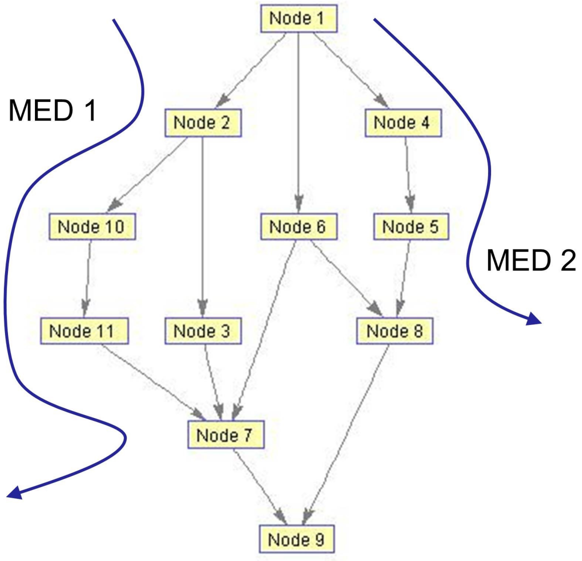


Fig. 13. Network topology with two MEDs

TABLE V
CALCULATION OF DISTANCE, TIME AND ENERGY COST FOR EVERY ROAD SEGMENT AND ROUTE FOR THE 2ND SIMULATION SCENARIO (FIGURE 13).

Road ID	*Dist (Km)*	*Time (min)*	*Vel (Km/h)*	*E cost (kWh)*	*E ind (kWh)*	*Total E (kWh)*
12	8	17	28.24	1.315	0.00	1.315
14	8	24	20.00	1.315	1.200	0.114
16	20	67	17.91	3.286	0.000	3.286
23	12	17	42.35	1.972	0.000	1.972
37	12	17	42.35	1.972	0.000	1.972
45	12	33	21.82	1.972	1.800	0.172
58	16	33	29.09	2.629	1.200	1.429
67	12	25	28.00	1.972	0.000	1.972
68	8	30	16.00	1.314	0.000	1.314
79	4	17	14.12	0.657	0.000	0.657
89	6	33	10.91	0.986	0.000	0.986
210	16	40	24.00	2.629	2.400	2.629
1011	16	40	24.00	2.629	2.400	2.629
117	16	40	24.00	2.629	2.400	2.629
Route						
1	*36*	**68**		*5.920*		*5.920*
2	*36*	*109*		*5.920*		*5.920*
3	**34**	*130*		*5.587*		*5.587*
4	*42*	*123*		*6.920*	4.200	**2.702**
5	*60*	*154*		*9.202*	7.200	**2.002**

the simulation area with a random initial energy between *4 and 24kW*. Vehicles with low initial energy book a free charging place in any of the two $MEDs$ and hence, manage to lower the consumed energy in the area.

Vehicles that choose to recharge during their travel in the simulation area need to wait for the MED at initial node where the recharging phase can begin. Bearing in mind that each MED is following a different route that includes road segments outside of the area and the current position of the MED at the time the booking application happens, the vehicle calculates the waiting time and decides whether to wait or not for the MED. In this scenario, vehicles do not have any time threshold, and wait no matter how long the time may be as long as the MED has a free charging place. This additional waiting time is represented in Figure 15.

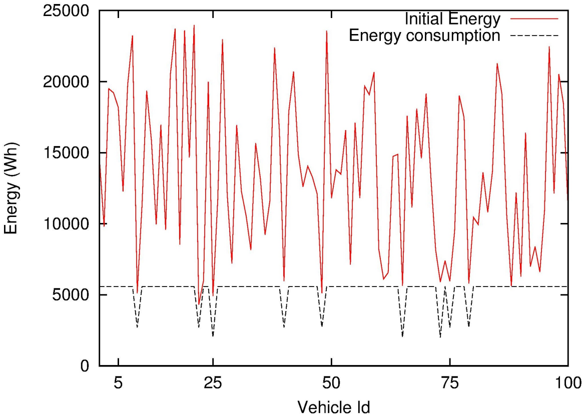


Fig. 14. Energy consumption of vehicles

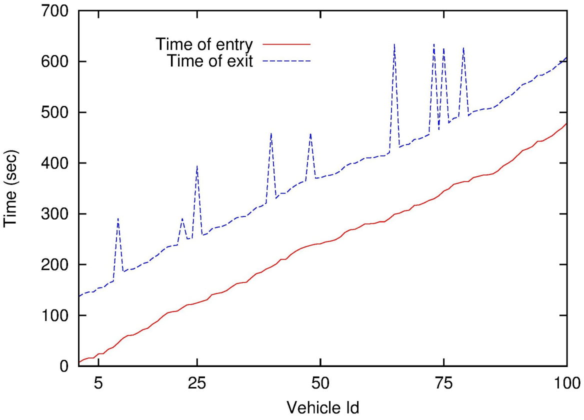


Fig. 15. Travel time of vehicles

When vehicle V_i follows any of the two $MEDs$ in order to recharge, in addition to more time, greater distance is traveled. The additional distance varies according to the MED that the vehicle chooses to follow, which is represented in Figure 16.

2) Impact of E threshold: Increasing the energy threshold causes an increase in demand for recharging from starving vehicles. However with the addition of a second MED in this situation, although leads to an increase in vehicles demanding to be recharged, this can be covered. Nevertheless, the system still has a maximum capacity based on the availability of the MED. If a MED has an empty slot, then V_i books it and waits for the MED to reach node *1* in order to begin recharging (*Route 4 or Route 5*). In the cases where the initial energy of the vehicle is above the threshold or no free charging place exists, it chooses to follow the shortest path in terms of distance (*Route 3*).

3) Density of vehicles: Changing the density of vehicles that move through the simulated area affects the performance of the system. For instance, lowering the density makes it

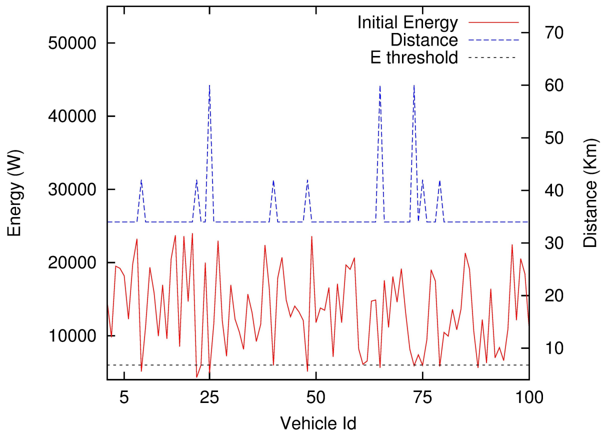


Fig. 16. Distance covered by every vehicle in the 2nd simulation scenario

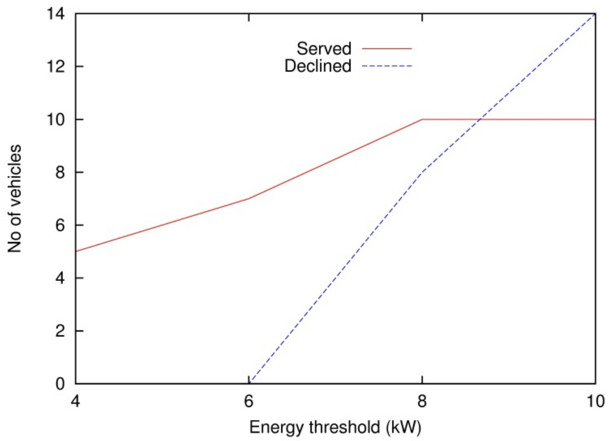


Fig. 17. System with 2 $MEDs$ can serve an excessive number of requests

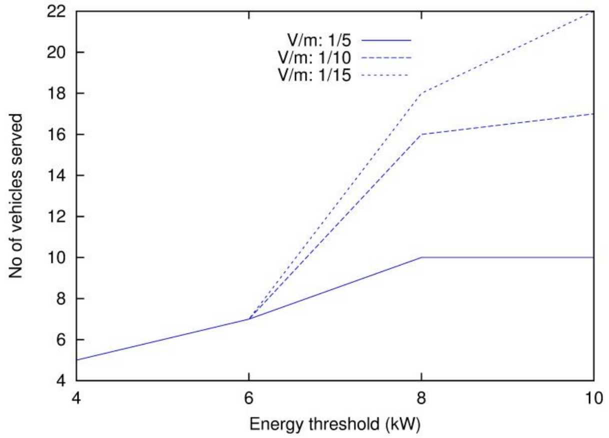


Fig. 18. Vehicles density affects system performance

TABLE VI
EVALUATION PERFORMANCE BASED ON REMAINING ENERGY E OF VEHICLES WHEN REACHING THE TARGET NODE 9.

Method	*No of Vehicles with with $E_{out} < 0$*	*Mean E_{out} kWh*	*Additional range %*
Energy path	0	10.186	
Shortest path	4	9.07211	-11.8 %
Fastest path	9	7.044	-44.03 %

capable of dealing with more demands for charging and hence, efficiency of the system is better for such scenarios (see Figure 18).

4) Impact on vehicles total range: In the shortest path and fastest time scenarios the number of vehicles that run out of energy and have to stop before reaching the final destination is given in Table VI. In the same table, the mean value of the remaining energy is also shown. It is evident that the energy efficient scenario, based on the use of $MEDs$, gives better outcomes and balances the consumed energy of vehicles. These values are calculated for the default values of the system (E_{th} *=6W, Vehicle rate =1/5*). Increasing the E_{th}, or lowering the rate that vehicles are injected into the system, the improvement in the remaining energy of the vehicles is even greater.

C. Intelligent MEDs

An important aspect of cooperative mechanisms is increased awareness about the conditions in the neighborhood of a vehicle. $MEDs$ receive several charging requests from vehicles and respond with regards to availability on a First Come – First Served basis without taking into account the real need of the vehicle requesting a booked place. This may lead to situations where vehicles with an immediate need for energy are unable to find a free booking place.

In order to cope with this problem, we modeled an enhanced mechanism where $MEDs$ decide to serve incoming requests based on the estimated additional energy that the vehicle will need in order to reach its destination. Vehicles, along with their request for energy, they also send the current energy level and their final destination. $MEDs$, using GPS and information about traffic on roads, can calculate the additional induced energy that each vehicle will need in order to reach its destination. . Based on this information, each MED can assess the level of importance of the requests, and serve those with greatest need, while declining to assist those with no real immediate requirements. The behaviour of the enhanced system is evaluated using the network with the two $MEDs$. All vehicles that send a request for recharging, also send information about their current energy level and final destination. With this information each MED decides whether to serve or decline the request.

In Figure 19, the impact of the cooperative mechanism on the systems performance is represented. In order to depict the differences in systems performance we inject vehicles at a lower rate *(1/10)* and we set E_{th} equal to *10kW*. The system now receives more requests for charging, while on the same time being able to meet most of them. The difference in the behaviour between the stand-alone and the cooperative mechanism lies in the appropriate selection of requests to be serviced.

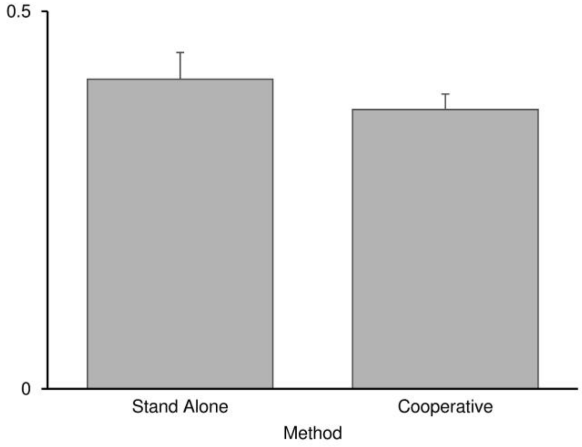


Fig. 19. Variance and standard deviation of parameter E_c

In order to represent this difference we calculate the mean and standard deviation of parameter *Ec*, which is calculated using Equation 12. The calculated metric represents the proportion of energy consumed to the initial energy of the vehicle. We can observe from Figure 19 that with the cooperative mechanism both mean value and standard deviation are decreased.

$$E_c = \frac{Energy\ consumption}{Initial\ Energy} \tag{12}$$

VII. Conclusions

In this article, we have demonstrated how mobile energy disseminators (MED) can facilitate EVs to extend their range in a typical urban scenario. Vehicles, based on several parameters, like time, energy and distance choose to follow longer but energy efficient paths. Making use of inductive charging $MEDs$ can act as mobile charging stations, thus improving the overall energy consumption of a fleet of vehicles. This improvement comes with a cost in time and distance traveled, but starving vehicles otherwise would have to stop or make longer re-routes in order to find a stationary station and recharge their batteries. Combining modern communications between vehicles and state of the art technologies on energy transfer, vehicles can extend their travel time without the need for large batteries or extremely costly infrastructure. Preliminary simulations show that applying some form of intelligence in how $MEDs$ take decisions about accepting or rejecting charging requests, further improves the performance of the method.

Our proposed method exploits IVC communications in order to eco-route electric vehicles taking advantage of the existence of $MEDs$. In the future, more complex scenarios are going to be investigated where vehicles will have more charging options to choose from, both stationary and mobile. The combined use of LTE and $DSCR$ capabilities is also going to be investigated, where vehicles will have the opportunity to communicate with a fleet of $MEDs$ and take more appropriate decisions based on less local information.

References

[1] T. R. Hawkins, B. Singh, G. Majeau-Bettez, and A. H. Strmman, “Comparative environmental life cycle assessment of conventional and electric vehicles,” *Journal of Industrial Ecology*, vol. 17, no. 1.

[2] H. de Wilde and P. Kroon, “Policy options to reduce passenger cars co2 emissions after 2020,” 2013.

[3] S. Lukic and Z. Pantic, “Cutting the cord: Static and dynamic inductive wireless charging of electric vehicles,” in *Electrification Magazine, IEEE 2013, vol. 1, no. 1. pp. 57-64*.

[4] N. Machiels, N. Leemput, F. Geth, J. Van Roy, J. Buscher, and J. Driesen, “Design criteria for electric vehicle fast charge infrastructure based on flemish mobility behavior,” 2013.

[5] G. Putrus, P. Suwanapingkarl, D. Johnston, E. Bentley, and M. Narayana, “Impact of electric vehicles on power distribution networks,” in *Vehicle Power and Propulsion Conference, 2009. VPPC ’09. IEEE*, 2009.

[6] A. R. Mario, A. A. Fadi, M. Gagnaire, and Y. Lascaux, “Telewatt: An innovative electric vehicle charging infrastructure over public lighting systems,” in *Proceedings of the 2nd International Conference on Connected Vehicles and Expo (ICCVE), Las Vegas, USA, December*, 2013.

[7] G. Jung, B. Song, S. Shin, S. Lee, J. Shin, Y. Kim, C. Lee, and S. Jung, “Wireless charging system for on-line electric bus(oleb) with series-connected road-embedded segment,” in *Environment and Electrical Engineering (EEEIC), 2013 12th International Conference on*, 2013.

[8] L. A. Maglaras, F. V. Topalis, and A. L. Maglaras, “Cooperative approaches for dymanic wireless charging of electric vehicles in a smart city,” in *Energy Conference (ENERGYCON), 2014 IEEE International*, May 2014, pp. 1365–1369.

[9] K. Boriboonsomsin, M. J. Barth, W. Zhu, and A. Vu, “Eco-routing navigation system based on multisource historical and real-time traffic information,” 2012.

[10] L. A. Maglaras, P. Basaras, and D. Katsaros, “Exploiting vehicular communications for reducing co2 emissions in urban environments,” in *in proceedings of the IEEE International Conference on Connected Vehicles (ICCVE 2013), Las Vegas, USA, December 2-6*, 2013.

[11] L. Zi-fa, Z. Wei, J. Xing, and L. Ke, “Optimal planning of charging station for electric vehicle based on particle swarm optimization,” in *Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2012.

[12] M. Badawy, N. Arafat, S. Anwar, A. Ahmed, Y. Sozer, and P. Yi, “Design and implementation of a 75 kw mobile charging system for electric vehicles,” in *Energy Conversion Congress and Exposition (ECCE), 2013 IEEE*, 2013.

[13] S. Deilami, A. Masoum, P. Moses, and M. A. S. Masoum, “Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile,” *Smart Grid, IEEE Transactions on*, 2011.

[14] K. Mets, T. Verschueren, W. Haerick, C. Develder, and F. De Turck, “Optimizing smart energy control strategies for plug-in hybrid electric vehicle charging,” in *Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP*, 2010.

[15] S. Bohn, M. Agsten, O. Waldhorst, A. Mitschele-Thiel, D. Westermann, and P. Bretschneider, “An ict architecture for managed charging of electric vehicles in smart grid environments,” in *Journal of Engineering*, 2013.

[16] D. Promiti, “Coordinating rendezvous points for inductive power transfer between electric vehicles to increase effective driving distance,” in *Proceedings of the 2nd International Conference on Connected Vehicles and Expo (ICCVE), Las Vegas, USA, December*, 2013.

[17] C. Donna, M. K. Kara, and K. Moby, “The electric vehicle charging station location problem: A parking-based assignment method for seattle,” in *Proceedings of the 92nd Annual Meeting of the Transportation Research Board*, 2013.

[18] I. Frade, A. Ribeiro, G. Goncalves, and A. Antunes, “Optimal location of charging stations for electric vehicles in a neighborhood in lisbon, portugal,” in *Transportation Research Record: Journal of the Transportation Research Board, No. 2252: 91-98*, 2011.

[19] M. R. Sarker, H. Pandzic, and M. A. Ortega-Vazquez, “Electric vehicle battery swapping station: Business case and optimization model.”

[20] M. Neaimeh, G. Hill, Y. Hubner, and P. Blythe, “Routing systems to extend the driving range of electric vehicles,” *Intelligent Transport Systems, IET*, vol. 7, no. 3, pp. 327–336, 2013.

[21] L. A. Maglaras and D. Katsaros, “Distributed clustering in vehicular networks,” in *Wireless and Mobile Computing, Networking and Com-*

munications (WiMob), 2012 IEEE 8th International Conference on, Oct 2012.

[22] N. Lagraa, M. B. Yagoubi, S. Benkouider *et al.*, "Localization technique in vanets using clustering (lvc)," *IJCSI*, 2010.

[23] M. Segata, F. Dressler, R. Lo Cigno, and M. Gerla, "A simulation tool for automated platooning in mixed highway scenarios," in *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 2012, pp. 389–392.

[24] L. A. Maglaras, J. Jiang, F. V. Topalis, and A. L. Maglaras, "Mobile energy disseminators increase electrical vehicles range in smart city," in *Hybrid and Electric Vehicle Conference, IET*, November 2014.

Energy consumption model over parallel programs implemented on multicore architectures

Ricardo Isidro-Ramírez
Instituto Politécnico Nacional
SEPI-ESCOM
M´exico, D.F.

Amilcar Meneses Viveros
Departamento de Computación
CINVESTAV-IPN
M´exico D.F.

Erika Hernández Rubio
Instituto Politécnico Nacional
SEPI-ESCOM
M´exico D.F.

Abstract—**In High Performance Computing, energy consumption is becoming an important aspect to consider. Due to the high costs that represent energy production in all countries it holds an important role and it seek to find ways to save energy. It is reflected in some efforts to reduce the energy requirements of hardware components and applications. Some options have been appearing in order to scale down energy use and, consequently, scale up energy efficiency. One of these strategies is the multithread programming paradigm, whose purpose is to produce parallel programs able to use the full amount of computing resources available in a microprocessor. That energy saving strategy focuses on efficient use of multicore processors that are found in various computing devices, like mobile devices. Actually, as a growing trend, multicore processors are found as part of various specific purpose computers since 2003, from High Performance Computing servers to mobile devices. However, it is not clear how multiprogramming affects energy efficiency. This paper presents an analysis of different types of multicore-based architectures used in computing, and then a valid model is presented. Based on Amdahl's Law, a model that considers different scenarios of energy use in multicore architectures it is proposed. Some interesting results were found from experiments with the developed algorithm, that it was execute of a parallel and sequential way. A lower limit of energy consumption was found in a type of multicore architecture and this behavior was observed experimentally.**

Keywords—Energy Consumption; Multicore Processors; Parallel Programs; Amdahl's law

I. Introduction

With the emerging of High Performance Computing, it was possible to carry out sophisticated calculations that before used to take a far longer time to be accomplished. This kind of computing has permitted to generate processes with much more elaborated and complex operations. Usually, HPC servers need high quantities of energy per unit time to work, for example, to implement a machine which execute a billion operations per second, or, in others words, a machine that would reach an exaflop, would require a big energy budget, comparable with the energy requirements of a town. Hence, energy consumption in data centers and supercomputers, is a topic which before did not have any importance, but is currently becoming more important every year.

Since 2004, multicore processors have been the object of interest of researchers that aim to exploit the potential of a die with more than one processing unit [1]. Clearly this is not only a case of study for servers and mainframes, because it is common to have multicore processors in mobile devices. This is an indicator that parallel programs are being used in various computing devices, from servers to smartphones.

A feature that makes a difference between mobile devices and other computer appliances, is that they are totally energy-bounded by a battery [2]. Across the years, battery technologies that feed devices have permitted to increase the capacity of these batteries without changing the size or weight of these components. However, mobile software developers often design and write applications which consider that the energy supplied is ideal, the battery is completely charged, and that usually it is not important to be aware of the energy impact caused by their applications. Neither are they worried enough by the fact that more hardware components being used and additional wireless requests will have a significant impact on the consumption of battery power.

A feasible solution for energy limitations is the introduction of parallel programming techniques that use efficiently all the computing resources available into modern smartphones. Actually, all industry is in fact using processors produced by companies such as Samsung, Qualcomm, Intel, etc. All these companies have been changing their hardware paradigms to offer more computing power into a single die; that includes the addition of independent processor chips joined by different mechanisms like shared cache, bridges, etc. In the market, the minimum number of cores present into a mobile multiprocessor is two. The reasons for this change are the constraints found in the traditional single-core paradigm, such as incremental heat and power dissipation [1].

The chipset landscape has a variety of multiprocessors divided by diverse features. Commonly, the designs are separated by their hardware architecture into symmetric and asymmetric [3], [4], but some researchers have also found another classification regarding their energy use features, e.g. a machine where processors can be turned off individually, or another machine where that is not possible [5].

Hardware classification separates processors by the computing capacity of their different cores. A sign of this class of processors is the presence of digital signal processors, GPU's and CPU's working together inside a chip. When cores have different capabilities, our multiprocessor is named heterogeneous or asymmetric. Otherwise, when all cores are similar, the multiprocessor is classified as homogeneous or symmetric.

We can distinguish three types of multicore processors if we consider its energy behavior. The first is a multicore where all cores are turned on, independently of how many cores a task needs. When the process is finished, all the cores will be turned off. This behavior implies that the same amount of energy is required to operate one or more cores in the processor. The second case is observed when the energy spent is proportional to the number of active cores; hence if a process uses one core the processor consumes a unit *n* of energy, if process uses two cores, it consumes *2n* units of energy, and so on. In the third case the processor has two operational voltages, one per each core, and another that feeds the entire processor. When all cores are turned off, each one of them operates in a energy state called *idle*, in addition a base voltage that feeds the other components of the processor is also present. Those different architectures are widespread in different appliances. In the traditional computing domain, symmetric architectures clearly dominate most of the market. But in mobile appliances, thanks to their low-power advantages that they offer [6], asymmetric chips are widely installed on them. In fact, asymmetric processors are even being used in server and mainframe environments due to these energy efficiency features [7].

Several authors have proposed models of energy consumption for multicore architectures. However, the experimental results have not matched the models, ie, the analysis of the models show energy savings [9], [11], [12], while in some experiments shown that programs on multicore architectures spend more energy when working with multiple threads [8], [15]. We believe that this contradiction is due to the models do not consider the types of processors, concerning the use of energy. This paper presents an analysis of different types of multicore-based architectures used in computing, and then a valid model is presented. Based on Amdahl's Law, a model that considers different scenarios of energy use in multicore architectures it is proposed.

This paper is organized as follows: Section II contains a brief review of some papers that have studied energy and power consumption in multiprocessors. Some of these papers deal with an adequate description of how power is used depending of many factors. Section III presents a description model based on the power using the muticore processors. The model has three stages, the last stage is the general, though the first two we consider necessary to understand the behavior of the model. Section IV presents some experimental results. These experiments are running the benchmark Linkpack in multicore processors. Finally, in Section V the conclusions of this work are presented.

II. Related work

Research over energy enhancements in multicore architectures, particularly about Amdahl's Law extensions, is focused on finding energetically sustainable architectures , using options like a set of rules that conforms a framework [5], or using techniques of CPU management, such as core offlining [8] and finding voltage and frequency optimal operating values of a multicore system [9].

Amdahl's Law has been used by some authors to model performance in computing scenarios [10]. Using their ideas to decide if multithreading programming gives performance and energy benefits, some of them get optimistic results [9], [11], [12], and others find pesimistic results when an enhanced energy consumption by adding threads is expected [8]. A discussion is centered in the fact of how many cores are needed to work properly, having in mind that more cores working in parallel represent execution speed, but require more energy to function. We pretend to use our model to describe energy consumption only in symmetric computers for the moment. This will be the base for our next study that we shall glimpse as future work.

Cho and Melhem [5] studied the mutual effects of parallelization, program performance, and energy consumption. Their proposed model was tested on a machine that could apply core offlining. With this work, they predict that more cores combined with a high percentage of parallel code in a process, helps to reduce energy use.

Basmadjian and de Meer [6] worked in the design of a software-based model of power consumption for multicores, and they suggested that it is important to bear in mind that the presence of more that one processing unit, has a direct relation in power behavior, because working with multiple cores implies computing sharing. They also mentioned that a hardware-level measure is not trivial, when the presence of a high quantity of circuits inside every multiprocessor is considered.

Hill and Marty [11] studied Amdahl's Law impact over various processors chip distributions, using homogeneous and heterogeneous dies. They suggested the idea of designing a chip that gives priority to global chip performance over individual core efficience. Even if a chip is composed of several cores, some of these could work together in order to offer a higher sequential performance; they concluded that an asymmetric multicore has better performance results against a symmetric multicore.

Sun and Chen [13] showed that multicore architectures are not limited properly by Amdahl's predictions, but a real drawback would be the disparity of technology improvements between CPU speed and memory latency. Thus, they conclude that there are not significant limitations in scalability in number of cores, but more research is needed to overcome memory limitations.

About using benchmark, Oi and Niboshi [14] made a study using two different CPUs and platforms. They did, in particular, a power measure for individual instruction and complete workload level. They studied the power behavior of a server, varying the operating clock frequency. The other interesting area that they tackled was performance, reflected in parameters like transaction response time and so on.

Isidro, Meneses and Hernández [15] showed previously to this paper that an energy-study design depends of many more factors such as hardware components that participate in application execution. They also proved that is possible to model and predict energy usage for an application, knowing the quantity of parallel and sequential code portions that a program has. This paper is process-oriented, as opposed to other studies which give more emphasis to hardware-oriented strategies. Finally, in this paper, we explain that there exist two different multicore architectures and three different energy usage models in multicore. Based on this affirmation, from Section 3 to Section 6, mathematical models of each one are explained.

III. Energy consumption model

In this paper the model of energy consumption for multicore processes is proposed. This model is an extension of Amdahl's law considering the types of processors for its energy behavior. It is well known that the relationship between power (Watts) and energy (Joules) is $Joules = Watts \times Seconds$. So, the idea is to have the power model and combine it with Amdahl's time model.

From [10][16], the speedup of a program to solve a problem of size N over p processing units $\psi(N,p)$ is divide the time to solve the problem of size n in a processor $T(N,1)$, between the time the same problem in p processors $T(N,p)$.

$$\psi(N,p) = \frac{T(N,1)}{T(N,p)}.$$

Such that, the time to solve a problem of size n by a processor, $T(N,1)$ or simply $T(N)$, can be split into the inherently sequential part $\sigma(N)$ and potentially parallel part $\rho(N)$.

$$T(N,1) = \sigma(N) + \rho(N). \tag{1}$$

And, the time to solve a problem of size N over p processing units is represented by

$$T(N,p) = \sigma(N) + \frac{\rho(N)}{p} + \kappa(N,p), \tag{2}$$

where $\kappa(N,p)$ is the overhead obtained by dividing the potentially parallel part in p processing units.

Consider a multiprocessor with n cores. Then CPU power is represented by W_{CPU}, and measured in Watts. Independently of the power usage architecture of a processor, it is possible to represent the entire power consumption by the addition of three parameters: base power, power of active cores, and idle power for all cores. Thus, the power of a processor with N cores with active cores p is:

$$W_{CPU}(N,p) = W_{base} + pW_{active} + nW_{idle}, \tag{3}$$

where $p < n$. Depending on the distribution of the process resources into a multiprocessor, these parameters could be significant or not. For example, (4) represents power required by a sequential program in a processor with n cores.

$$W_{CPU}(N,1) = W_{base} + W_{active} + nW_{idle}. \tag{4}$$

Now, having time and power expressions, it is possible to model energy consumption in Joules(J). Depending if the program is sequential (one thread), or parallel with p threads, two cases are presented.

The energy required to solve a problem of size N on a single core, $J(N,1)$, can be split into the inherently sequential part $J_\sigma(N,1)$ and potentially parallel part $J_\rho(N,1)$:

$$J(N,1) = J_\sigma(N,1) + J_\rho(N,1). \tag{5}$$

The energy required for the sequential and parallel portions are modeled with (6) and (7) respectively:

$$\begin{aligned} J_\sigma(N,1) &= J_\sigma(N) = \sigma(N)\left(W_{base} + W_{active} + nW_{idle}\right), & (6)\\ J_\rho(N,1) &= J_\rho(N) = \rho(N)\left(W_{base} + W_{active} + nW_{idle}\right), & (7)\end{aligned}$$

Now, the energy required to solve a problem of size N on p cores in a processor with n cores, where $p < n$, is represented in (8):

$$J(N,p) = J_\sigma(N,1) + J_\rho(N,p) + J_\kappa(N,p). \tag{8}$$

The sequential part is the same that (6). The energy required to execute the parallel portion over p cores is:

$$J_\rho(N,p) = \frac{\rho(N)}{p}\left(W_{base} + pW_{active} + nW_{idle}\right). \tag{9}$$

The next step for the model is to represent the energy speedup, $\psi_J(N,p)$, that a parallel program will reach against their sequential version.

$$\psi_J(N,p) = \frac{J(N,1)}{J(N,p)} = \frac{J_\sigma(N) + J_\rho(N)}{J_\sigma(N) + J_\rho(N,p) + J_\kappa(N,p)} \tag{10}$$

We can consider three general scenarios for the power of a multicore processor. Equation (3) is a general schema of power consumption was presented, showing the three parts that generally compose the entire CPU power consumption. Now, specific formulae for each architecture will appear. The first scenario consists in a machine model where all cores are turned on, independently of the task size and number of threads used. In this case, CPU power uses only a base power value, and *idle* and *active* states are discarded, as show in (11). The second scenario works with just the necessary cores that a process needs, the rest of cores remain turned off. Hence, CPU power usage is represented by p times the *active* power of a unitary core, as show in (12). The last architecture is found when a processor has baseline power and also an individual operation power per core. In that case, the CPU power is the sum of base power, *idle* power and *active* power, as show in (13).

$$\begin{aligned} W_{CPU} &= W_{base}, & (11)\\ W_{CPU} &= pW_{active}, & (12)\\ W_{CPU} &= W_{base} + pW_{active} + nW_{idle}. & (13)\end{aligned}$$

A. Model applied to constant power usage multicore

The first scenario in our classification that we have observed is any multicore architecture where energy consumption is directly proportional to the number of cores that are turned on (from one to n), i.e. when the power using the multicore processor is constant. Its means that the W_{CPU} is described in (11)

$$\begin{aligned} \frac{J(N,1)}{J(N,p)} &= \frac{J_\sigma(N,1) + J_\rho(N,1)}{J_\sigma(N,1) + J_\rho(N,p) + J_\kappa(N,p)}\\ &\leq \frac{J_\sigma(N,1) + J_\rho(N,1)}{J_\sigma(N,1) + J_\rho(N,p)}\\ &\leq \frac{\sigma(N)W_c + \rho(N)W_c}{\sigma(N)W_c + \frac{\rho(N)}{p}W_c}\\ &\leq \frac{W_c(\sigma(N) + \rho(N))}{W_c(\sigma(N) + \frac{\rho(N)}{p})}\\ &\leq \frac{\sigma(N) + \rho(N)}{\sigma(N) + \frac{\rho(N)}{p}} & (14)\end{aligned}$$

If we consider f as the percentage of inherently sequential execution time [10][16]:

$$f = \frac{\sigma(N)}{\sigma(N) + \rho(N)}. \tag{15}$$

And if we substitute (15) in (14), then we obtain

$$\frac{J(N,1)}{J(N,p)} \leq \frac{1}{f + \frac{1-f}{p}}. \tag{16}$$

We can see that the resulting inequality (16) is Amdahl's Law. Where f represents the percentage of inherently sequential execution time [10][16].

For this case, the model of constant energy, and based on (3), it is only necessary a base level power for each core, since an *idle* power is non existent here. Equation (13) describes adequately this multicore behavior. Figure 1 represents the speedup curves depending of the quantity of sequential code f that a process contains. The graph illustrates that speedup increases as a combination of high potential parallel code and a large amount of cores. The acceleration shown in Figure 1 for $f < 0.2$ means that you can have energy savings when more cores are used in a multicore processor to solve a parallel process.

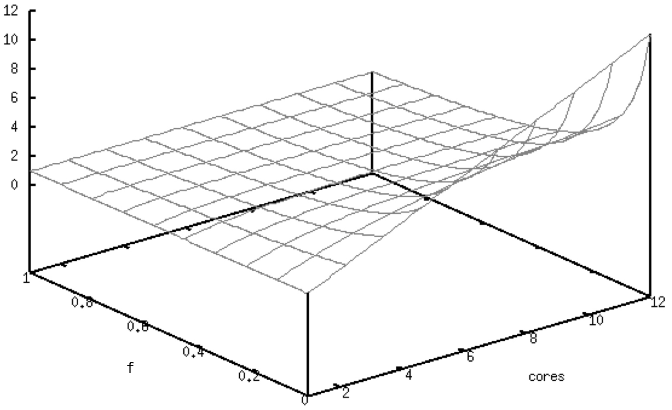


Fig. 1: Energy speedup for a constant energy usage, depending of the number of cores and the sequential portion f of the process

B. Model applied to offlining multicore

In the second scenario, it is not considered the base power or idle power, ie, power is proportional to the number of active cores. This scenario is an offlining energy architecture. If an offlining energy architecture is present in the energy architecture, then only the active power for each used core exists. The total power is modeled with (12), and then, the speedup is given by:

$$\begin{aligned}
\frac{J(N,1)}{J(N,p)} &= \frac{J_\sigma(N,1) + J_\rho(N,1)}{J_\sigma(N,1) + J_\rho(N,p) + J_\kappa(N,p)} \\
&\leq \frac{J_\sigma(N,1) + J_\rho(N,1)}{J_\sigma(N,1) + J_\rho(N,p)} \\
&\leq \frac{\sigma(N)W_{cpu}(1) + \rho(N)W_{cpu}(1)}{\sigma(N)W_{cpu}(1) + \frac{\rho(N)}{p}W_{cpu}(p)} \\
&\leq \frac{\sigma(N)W_{active} + \rho(N)W_{active}}{\sigma(N)W_{active} + \frac{\rho(N)}{p}pW_{active}} \\
&\leq 1.
\end{aligned}$$

Hence, the behavior of this scenario tells us that the lower bound for the energy used in a process with this kind of multicore, corresponds always with the energy used by a sequential execution, expressed in (18):

$$\psi_J(N,p) = \frac{J(N,1)}{J(N,p)} \leq 1, \tag{17}$$

$$J(N,1) \leq J(N,p). \tag{18}$$

Therefore, in this scenario no energy savings when parallel programs running on multicore processors. However, having a lower limit can be referenced to find the best energy performance of a parallel program.

C. General case: Model applied to multicores with base and individual core power

The third energy scenario corresponds to all multicores when there exists a base power that feeds the entire chip, and adds more power depending of the number of cores that a process requests. Additional to these power values, each core has an *idle* power that must be taken into account. Consider that all cores are symmetric for this model.

$$\begin{aligned}
\frac{J(N,1)}{J(N,p)} &= \frac{J_\sigma(N,1) + J_\rho(N,1)}{J_\sigma(N,1) + J_\rho(N,p) + J_\kappa(N,p)} \\
&\leq \frac{J_\sigma(N,1) + J_\rho(N,1)}{J_\sigma(N,1) + J_\rho(N,p)} \\
&\leq \frac{\sigma(N)W_{cpu}(1) + \rho(N)W_{cpu}(1)}{\sigma(N)W_{cpu}(1) + \frac{\rho(N)}{p}W_{cpu}(p)},
\end{aligned}$$

if $W_{CPU}(p) = W_{base} + pW_{active} + nW_{idle}$,

$$\Psi_J = \frac{J(N,1)}{J(N,p)} \leq \frac{1}{f + (1-f)(\frac{W_{cpu}(p)}{pW_{cpu}(1)})} \tag{19}$$

The energy speedup of this case depends on the power used in the execution of sequential process and power used in the execution of parallel processing, as show in (19).

Observe that depending on the values of power, the speedup graphics present variations. We can predict in an example with low values of base, idle and active power, a behavior like the one shown in Figures 2 and 3:

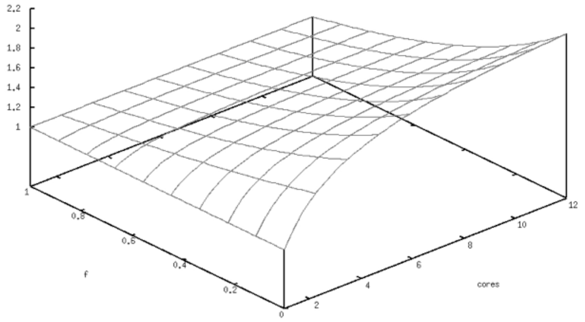


Fig. 2: Energy speedup for the third scenario, with $W_{active} = 10w$, $W_{base} = 1w$, $W_{idle} = 1w$ and n from 0 to 12

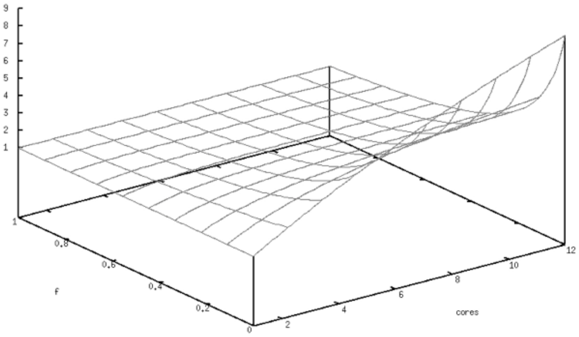


Fig. 3: Energy speedup for the third scenario, with $W_{active} = 2w$, $W_{base} = 40w$, $W_{idle} = 2w$ and n from 0 to 12

IV. Experimental results

To test the model, the benchmark Linpack was used to stress the multicore processor [17], [18]. Linpack is a measure of a computer's floating point rate of execution, that is determined by running a computer program that solves a dense system of linear equations. Despite the existence of other benchmarks that are focused on measuring diverse aspects of interest in software such as peak-use, complexity, cost per performance, among other aspects, Linpack was implemented because the issue of interest in this research is to stress the entire processor, including the floating-point unit.

In order to perform measurements, two devices were used to make the respective tests, these devices are listed below:

- Tablet Samsung Galaxy Tab 2 10.1, version GT-P5110, with Android OS 4.0 and 1.0 GHz OMAP4430 dual-core 45 nm ARM Cortex-A9.

This device correspond to mobile environment. However the device has multiple cores and it is possible to make an analysis based on execution time of an algorithm that could employ a variable quantity of resources, running in a parallel or a sequential mode. Also Linpack benchmark has been used in a server environment, but with few changes, it was programmed by us for a mobile scenario also.

Linpack in mobile version is able to execute vector-vector, vector-matrix and matrix-matrix products, and was developed under various multicore platforms. We made several implementations using POSIX threads and Android Java threads to generate parallel code. In this work, Linpack benchmark implementations were used over Android.

A. Working with Android

The tests were conducted in low-level Android libraries written in C and Java libraries used. The purpose of these comparisons is to detect the same energy behavior of the model regardless of whether the program is working layer, that is, if you work with programs that run directly on multicore or go through a virtual machine.

Inside every Android device there is a special implementation of VM named *Dalvik VM*, whose goal is to run Java applications on mobiles. The *Dalvik VM* is well suited for a portable scenario, offering a reduced energy consumption and a better performance, compared with Java Virtual Machine [19]. Other features of DVM that mobile devices are capable to use are low-memory requirements and delegation of tasks for the operating system, like memory and threads management.

However, we programmed Linpack, writing all the logic in Java Native Code or *pthreads* as a JNI module[1]. Remember that there are two C/C++ libraries, one of them is bionic, present in Android based systems, while the other set of functions is `glibc`, present in Linux based systems. The most profound characteristics that have permitted the use of bionic instead `glibc` in Android are the limited storage and a lower CPU speed. Actually, bionic is a lightweight library that comes with some bounds, compared with the GNU version.

Another issue to consider is that many operating systems are ruled by policies named CPU frequency schedulers, that scale the chip frequency up or down in order to save power. Operating frequency can be: Scaled automatically depending on the systems loads, chosen by the user, or managed in response to ACPI events. This point is crucial in mobile devices, because a daemon that puts our device in an idle state our device if unused is convenient. Some devices can even support more than one governor if it is permitted in a list defined by the manufacturer. The common feature in mobile governors is that usually they seek the least possible CPU usage and, as a consequence, they try to sleep the processor cores as much as possible.

Galaxy Tab 2 has implemented *Interactive* governor, other available governors for most of the Android devices are: *OnDemand, Performance, Userspace, Powersave, Conservative*, among others. The differences among them reside in the time lapse that one keeps working the multicore, and the performance requirements of the user or application executed. In this paper, we did not consider the effects of each one of these schedulers, and the experiments worked with the original governor installed by default.

The results of the experiments are shown next, where Linpack matrix products with one and two threads were executed. This operation has an $O(n^3)$ complexity, enough to obtain processor usage values near to 100% of processor use.

Both experiments were executed in the same platform. In order to compare energy behavior between Android JNI and pure Java, we prepared two applications that perform the same

[1] In this paper, we use JNI module as the code section that runs with *pthreads*

function, but were developed differently. The first application is an Android version that makes the heavier computing in a native language, and the second version does the entire job over Java.

Table I illustrates time and power required to execute the entire program, and the quantity of resources used for sequential and parallel sections of the program when it runs over one thread. Table II shows the same information for a program that runs in two parallel threads.

Problem size	Sequential section time	Parallel section time	Total time	Sequential section energy	Parallel section energy	Total energy
500x500	0.0173	10.649	10.66	0.067	6.207	6.275
1000x1000	0.0474	110.223	110.2	0.084	61.631	61.71
1500x1500	0.0848	427.094	427.1	0.078	236.12	236.2
2000x2000	0.1585	1065.523	1065.6	0.058	593.34	593.4

TABLE I: Time (in seconds) and energy (in Joules) values obtained with single precision matrix-matrix product, using JNI and Android with one thread.

Problem size	Sequential section time	Parallel section time	Total time	Sequential section energy	Parallel section energy	Total energy
500x500	0.0173	5.295	5.312	0.067	6.207	6.275
1000x1000	0.0484	55.737	55.78	0.085	61.854	61.94
1500x1500	0.0849	228.781	228.8	0.088	253.133	253.2
2000x2000	0.1298	611.147	611.27	0.144	670.552	670.6

TABLE II: Time (in seconds) and energy (in Joules) values obtained with single precision matrix-matrix product, using JNI and Android with two threads.

Table I and Table II show better performance when the program has been built in parallel. On the other hand, energy consumption has not the same effect, because the energy required grows if the program is multithreading.

The second experiment executes the same matrix-matrix algorithm but discarding the use of a native language. Hence, we only use Java to generate all the workload.

Problem size	Sequential section time	Parallel section time	Total time	Sequential section energy	Parallel section energy	Total energy
500x500	0.032	15.968	16.0	0.015	7.386	7.401
1000x1000	0.058	193.782	193.8	0.024	83.175	83.2
1500x1500	0.234	751.166	751.4	0.126	315.274	315.4
2000x2000	0.406	2028.374	2028.8	0.191	851.409	851.6

TABLE III: Time (in seconds) and energy (in Joules) values obtained with single precision matrix-matrix product, using pure Java with one thread.

Table III shows the execution time and energy used for matrix multiplication operations of different sizes, but using only one thread for all the work. Table IV shows also the time needed and energy consumed for the same program, but executed with two threads instead.

From Table I and III, we can see that the time and energy used to execute the task is lower for the JNI version of the program. Also, the potentially parallel section of the multiplication matrix algorithm is by far, the heaviest section of the program.

Problem size	Sequential section time	Parallel section time	Total time	Sequential section energy	Parallel section energy	Total energy
500x500	0.036	7.9824	8.018	0.016	7.501	7.517
1000x1000	0.056	98.04	98.09	0.022	86.048	86.07
1500x1500	0.201	400.959	401.1	0.137	337.563	337.7
2000x2000	0.407	1173.365	1173.7	0.203	962.097	962.3

TABLE IV: Time (in seconds) and energy (in Joules) values obtained with single precision matrix-matrix product, using pure Java with two threads.

From Tables II and IV, one may verify that parallelism gives benefits in both versions of the program, independently of the mechanism used to generate the code. Also, energy consumption for both programs is still lower when we use one thread to execute the program. Finally, while the performance gets speedups from 1.7x for 2000x2000 matrix product, to 2.0x for 500x500 matrix, we observe an opposite effect in energy consumption.

In Figure 4, observe that the least execution time is obtained with parallel applications, regardless if we use Java or JNI, and making a comparison among 4 program versions, a JNI implementation with 2 threads results the fastest of all.

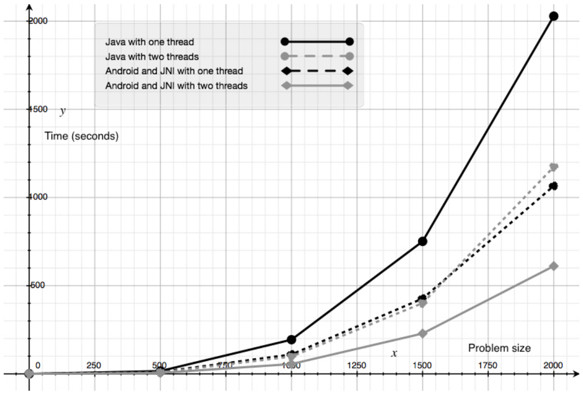


Fig. 4: Time values obtained with Linpack, using Java and JNI.

Figure 5 shows that both sequential programs (Java and JNI) produce energy savings, compared with parallel versions of the same program; this is in agreement with (18) about offlining multicores. Finally, both Figures highlight the gap between parallel and sequential results, which increases depending on the size of the problem.

The tool used to measure energy consumption and execution time was an application called PowerTutor [20], which estimates power usage by reading parameters available in Linux-based systems, like Android. For example, knowing the energy used by the processor is possible thanks to the parameters `/proc/stat` and `/proc/cpuinfo` that reveal the operating frequency of the chip and determine the power used during a task execution. Furthermore, one advantage of using this method to estimate power and energy usage resides in the difficulty to obtain power readings using a power meter in a very reduced environment. The other fact that supports this

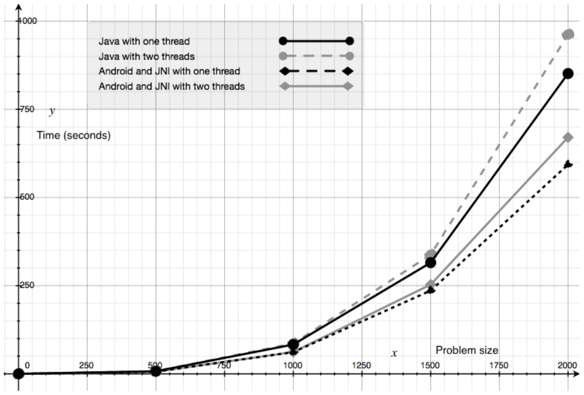


Fig. 5: Energy values obtained with Linpack, using Java and JNI.

method for our experiments is the low error rate of PowerTutor, ca. 6.27% [21].

V. Conclusion

One of the main contributions of this model, is the possibility of offering an explanation for the apparently contradictory results reported by several authors, because while some of them show energy savings using parallel programming in multicore processor [9], [11], [12], there is work which shows that energy consumption is punished by parallel techniques [8], [15].

Energy consumption contradictions exist because those models do not consider the existence of different power usage scenarios in multicore processors. It is very likely that the reason why authors finds energy savings with parallel applications is a processor that works with the first energy usage scenario exposed, where the multicore has a constant energy usage, independently of the number of turned on cores.

The model works for parallel programs running on multicore architectures. The model is tuned so that the number of processes that are used are less than or equal to the number of processor cores. However, with a number of threads higher than the number of cores n, it behaves as if working with n cores.

The model considers three main scenarios based on the power using the muticore processors. In the first scenario, where the processor uses the same power regardless of the number of active cores, Amdahl's law is obtained. Acceleration performance is proportional to the acceleration in energy consumption, which results in energy savings. In fact it is proposed in [6] that there is an energy benefit when computing resources are shared. In our model that behavior is reflected in the first scenario, where a constant power feeds the entire chip. This is to say that when a multicore processor has energy savings with a multithreading program, it is because all computing resources are shared in the processor.

In case where offlining scenario is present, we can observe the existence of a lower bound of energy consumption explained in this work, present when the sequential version of a program runs over the processor. When the number of threads of a parallel program increases, a reduction in execution time is obtained, while an increase in energy consumption will take place at the same time. Such behavior is seen in Figures 4 and 5, where we show that the lower bound in energy consumption corresponds to a totally sequential execution program. In this scenario no energy savings when parallel programs running on multicore processors. However, having a lower limit can be referenced to find the best energy performance of a parallel program.

In particular, in our experiment in android, the behavior of the lower bound of energy of the sequential version remains. With these experiments, we demonstrate that there is a notable improvement using native methods over Java, in performance and energy consumption, independently of the number of threads occupied. So, we claim that reducing virtual machine usage, gives a better performance in a high demanding computing application within the mobile scenario. We can infer that the behavior of power that has this processor is similar to the offlining scenario.

The third scenario is the most general and energy savings depend on the amount of resources that the multicore processor (19). The

$$\frac{W_{cpu}(p)}{pW_{cpu}(1)}$$

ratio that appears in (19), and CPU power (3) are the keys to our model to explain the behavior of parallel execution in different architectures of multicore processors. For example when $W_{cpu}(p) = W_{cpu}(1)$, general model (19) exhibits the behavior of the first scenario. And when $W_{base} + nW_{idle}$ is near to zero, general model (19) describes the behavior of the second scenario.

As future work, we will develop measurements in other computing environments, such as servers, laptops and desktop computers to verify the pattern of energy consumption.

Acknowledgment

R.I.R. thanks Section of Research and Graduate Studies (SEPI) of ESCOM-IPN, and Julieta Medina Garcia for scholarship support (SIP project 20131498). A.M.V. and E.H.R. thanks Cinvestav-IPN, and Section of Research and Graduate Studies (SEPI) of ESCOM-IPN, by the resources provided and the facilities for this work.

References

[1] S. H. Fuller and L. E. Miller, "Computing performance: Game over or next level?" *The National Academies Press*, pp. 31–38, 2011.

[2] P. Zheng and L. M. Ni, *Smart Phone and Next Generation Mobile Computing*. Elsevier Science and Tech, 2006.

[3] L. Carro and M. B. Rutzig, "Multi-core systems on chip," *Handbook of Signal Processing Systems*, pp. 485–514, 2010.

[4] D. H. Woo and H.-H. S. Lee, "Extending amdahl's law for energy-efficient computing in the many-core era," *Computer*, vol. 41, no. 12, pp. 24–31, 2008.

[5] S. Cho and R. G. Melhem, "On the interplay of parallelization, program performance, and energy consumption," in *IEEE Transactions On Parallel and Distributed Systems*, vol. 21. IEEE, 2010, pp. 342 – 353.

[6] R. Basmadjian and H. de Meer, "Evaluating and modeling power consumption of multi-core processors," in *Future Energy Systems: Where Energy, Computing and Communications Meet (e-energy)*, 2012, pp. 1–10.

[7] Z. Ou, B. Pang, Y. Deng, and J. Nurminen, “Energy- and cost-efficiency analysis of arm-based clusters,” in *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*. IEEE, 2012, pp. 115–123.

[8] A. Carroll and H. Heiser, “Mobile multicores: Use them or waste them,” *5th Workshop on Power-Aware Computing and Systems HotPower ’13*, 2013.

[9] S. M. Londoño and J. P. de Gyvez, “Extending amdahl’s law for energy-efficience,” in *Energy Aware Computing (ICEAC), 2010 International Conference*. IEEE, 2010, pp. 1–4.

[10] G. M. Amdahl, “Validity of the single-processor approach to achieve large-scale computing capabilities,” in *AFIPS Conference*, vol. 30, 1967, pp. 483–485.

[11] M. D. Hill and M. M.R, “Amdahl’s law in the multicore era,” *Computer*, vol. 41, no. 7, pp. 33–38, 2008.

[12] R. Swapnoneel, A. Rudra, and A. Verma, “An energy complexity model for algorithms,” in *4th Conference on Innovations in Theoretical Computer Science ITCS ’13*, 2013, pp. 283–304.

[13] X.-H. Sun and Y. Chen, “Reevaluating amdahl’s law in the multicore era,” *Journal of Parallel Distributed Computing*, vol. 70, no. 2, pp. 183–188, 2010.

[14] H. Oi and S. Niboshi, “Power-efficiency study using specjenterprise2010,” in *Systems Conference (SysCon), 2013 IEEE International*, 2013, pp. 812–817.

[15] R. I. Ramirez, E. H. Rubio, and A. M. Viveros, “Energy consumption in mobile computing,” in *Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference*, 2013, pp. 132–137.

[16] M. J. Quinn, *Parallel Programming in C with MPI and OpenMP*. McGraw-Hill Education Group, 2003.

[17] J. Dongarra, P. Luszczek, and A. Petitet, “The linpack benchmark; past, present and future,” in *Concurrency and Computation: Practice and Experience, 15*, 2003, pp. 803–820.

[18] J. Dongarra, “Performance of various computers using standard linear equation software,” in *SIGARCH Computer Architecture News*. ACM, 1992, pp. 22–44.

[19] K. Paul and A. Kumar, “Android on mobile devices: An energy perspective,” in *In Computer and Information Technology (CIT), 2010*, 2010, pp. 2421–2426.

[20] “Power tutor: A power monitor for android-based mobile platforms.”

[21] L. Z. Lide Zhang, B. S. Tiwana, R. P. Dick, and Z. Qian, “Accurate online power estimation and automatic battery behavior based power model generation for smartphones,” in *Hardware/Software Codesign and System Synthesis (CODES+ISSS*. IEEE, 2010, pp. 105–114.

[22] iOS Developer Series, *Threading Programming Guide*, Apple Inc., 2014.

Graph-based Semi-Supervised Regression and Its Extensions

Xinlu Guo
Graduate School of System Informatics
Kobe University
Kobe, Japan 657–8501

Kuniaki Uehara
Graduate School of System Informatics
Kobe University
Kobe, Japan 657–8501

Abstract—**In this paper we present a graph-based semi-supervised method for solving regression problem. In our method, we first build an adjacent graph on all labeled and unlabeled data, and then incorporate the graph prior with the standard Gaussian process prior to infer the training model and prediction distribution for semi-supervised Gaussian process regression. Additionally, to further boost the learning performance, we employ a feedback algorithm to pick up the helpful prediction of unlabeled data for feeding back and re-training the model iteratively. Furthermore, we extend our semi-supervised method to a clustering regression framework to solve the computational problem of Gaussian process. Experimental results show that our work achieves encouraging results.**

Keywords*—*Semi-supervised learning; Graph-Laplacian; Regression; Gaussian Process; Feedback; Clustering

I. Introduction

Regression is a fundamental task in data mining and statistical analysis. A regression task aims to analyze and model the relationship between variables so that the value of a given variable can be predicted from one or more other labeled variables. By using enough labeled training data, supervised regression algorithm can learn reasonably accurate model. However, in many machine learning domains, such as bioinformatics and text processing, labeled data is often difficult, expensive and time consuming to obtain. Meanwhile unlabeled data may be relatively easy to collect in practice. For this reason, in recent years, semi-supervised regression has received considerable attention in the machine learning literature due to its potential in utilizing unlabeled data to improve the predictive accuracy [6] [28].

An early semi-supervised regression method is iterative labeling [9], such as co-training algorithm [4][27], which employs supervised regressors as the base learners, then labels and selects unlabeled data in an iterative process. Similarly [5] performed another co-training style semi-supervised regression algorithm by employing multiple learners. Although these methods achieved considerable improvements, they didn't take full advantage of the inherent structure between labeled and unlabeled data. Indeed, they just kept the supervised learning algorithm and changed the form of the labels of data, i.e., they label and relabel the unlabeled data iteratively. Unfortunately, the iterative process causes computational problems for large datasets.

Besides co-training, regularization based method has also been widely employed in the semi-supervised regression [11][15][23][3]. This method combines a regularization term of all labeled and unlabeled data, with the predictive error of labeled data into a criterion. In such a criterion, the unlabeled data can help to get a better knowledge for what parts of the input space that the predictive function varies smoothly in. A variety of approaches using the regularization term have been proposed. Some well-known regularization terms are graph Laplacian regularizer [29], Hessian regularizer [10], parallel field regularizer [14], and so on. These methods have enjoyed a great success. However, they are transductive, which means they only work on the observed labeled and unlabeled training data and can't handle the unseen data.

In this paper, we propose an inductive semi-supervised regression model through incorporating graph prior information into the standard Gaussian process (GP) regression. Our method firstly builds an adjacent graph over all the labeled and unlabeled data. Then we consider the adjacent graph as a prior and incorporate it with the standard GP prior to generate a new GP prior condition on the graph and a graph-based covariance function. From the new conditional prior and the graph-based covariance function, the marginal likelihood and the prediction distribution of semi-supervised GP regression are derived. Since the prediction from the GP model takes the form of a full predictive distribution, the unseen data can also be predicted easily.

Additionally, to further boost the learning performance, we also extend our semi-supervised method to a feedback framework. The early semi-supervised learning methods, such as self-learning [19] and co-training [27], usually make use of a supervised learning algorithm to label and select unlabeled data in an iterative process. And these methods have been proved to be effective in improving the prediction accuracy. Thus, the predictions of the learning process must contain some valuable information, and under some metrics, they can help to construct more accurate model. In other words, when a learning process is performed repeatedly, we gain extra information from a new source: past unlabeled examples and their predictions, which can be viewed as a kind of experience. This kind of experience serves as a new source of knowledge related to the prediction model. The new knowledge provides the possibility of improving the performance of our semi-supervised GP regression. In this paper, to take advantage of such extra information, we also employ a feedback algorithm to pick up the helpful prediction of unlabeled data for feeding back and re-training the model iteratively.

Furthermore, we empirically demonstrate a further extension of the semi-supervised GPr. GP has the computational problem due to an unfavorable cube scaling ($O(N3)$) during training, where, N is the number of training data. In recent years, many methods have been proposed to address this problem: sparse GP approximation [24][20][12], localized regression [7][17]. In our work, we describe a clustering regression framework in order to bring the scaling down. Specifically, a clustering algorithm is employed as the first step in the process for identifying regions that have similar characteristics. Then for each cluster, a local semi-supervised regression model is built to describe the relationship between inputs and outputs. By partitioning the dataset and learning models locally, the computational cost for each local model is cubic only in the numbers of data points in each cluster, rather than in the entire dataset. As a result, even for large dataset, it can lead to a more favorable training scaling.

This paper is organized as follows: In Section 2, we discuss some related work. In Section 3 we give some preliminaries and a brief overview of the Gaussian process regression. The problem statement and our main theorem, as well as the key models are detailed in Section 4. In Section 5, we lay out an extension algorithm that detects usefull predictions and feeds them into the training set. In Section 6 we experimentally compare our method with the state-of-art approaches and make a detailed discussion. According to the results, we find out a problem of our method, and also describe a clustering regression framework to fixing it. Finally, section 7 concludes our work.

II. Related Work

Our work is closely related to several semi-supervised learning methods. One is the semi-supervised classification method proposed by [21]. We both define a prior for the graph variables and attempt to incorporate it into the standard GP probability framework to derive a posterior distribution of latent variables condition on the graph. However, all their derivations are focused on semi-supervised classification problem but not the regression problem. Thus, we will not discuss in more detail.

Additionally, our work is also similar to Zhang's method of semi-supervised multi-task regression (SSMTR)[26]. It seems that we both construct an adjacent graph and incorporate the prior of this graph with the GP prior to generate a semi-supervised data dependent kernel function that defines over the entire data space. But there are several differences in deed. In Zhang's paper, they proposed a new GP likelihood $\prod_{i=1}^{m} p(y^i|X^i)p(\theta^i)$ for the supervised multi-tasks regression (named SMTR), and then changed the kernel function of the model to the semi-supervised kernel function, to extend their model to the semi-supervised setting. Here the semi-supervised kernel function has also been used in classification task before [22]. Actually, the prediction formulation of SMTR $p(y_*^i|x_*^i, X^i, y^i)$ is the same as standard GP but with different kernel functions. In our paper, we don't simply change the kernel function in a supervised GP, but take advantage of the prior of the adjacency graph to derive a new likelihood condition on the graph $p(y|X, \mathcal{G})$ and a conditional prediction distribution $p(y_*|x_*, X, y, \mathcal{G})$, which is the training model and prediction model for the semi-supervised regression. In other words, the major difference between our method and SSMTR is that the training and prediction models are totally different. Moreover, in Zhang's method, because of the large number of parameters, it is difficult to estimate the optimal values simultaneously. So the parameters are optimized through an alternating optimization algorithm. However, the parameters in our work are estimated by using the gradient descent method to minimize the negative log conditional likelihood $p(y|X, \mathcal{G})$, which means that the training processes of two methods are different.

III. An Overview of Gaussian Process Regression

GP has been proved to be a powerful tool for the purpose of regression. The important advantage of GP is the explicit probabilistic formulation. This not only provides probabilistic predictions but also gives the ability to infer model parameters. Here, we offers a brief summary of GP for supervised regression, see [18] for more details. We assume that the input training data is given as $X_D = \{X_L, X_U\} = \{x_1, \ldots, x_\ell, x_{\ell+1}, \ldots, x_N\}$, where $x_i \in R^d$, N is the total number of input data and ℓ is the number of labeled data. X_L and X_U denote the inputs of labeled and unlabeled dataset respectively. We use $y = \{y_1, \ldots, y_\ell\}$ to represent the corresponding outputs of labeled data X_L.

In supervised GP regression, the corresponding output label y is assumed relating to an latent function $f(x)$ through a Gaussian noise model: $y = f(x) + \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(m, c)$ is a Gaussian distribution with mean m and covariance c. The regression task is to learn a specific mapping function $f(x)$, which maps an input vector to a label value. Usually, a zero-mean multivariate Gaussian prior distribution is placed over f. That is:

$$\begin{aligned} p(f|X_L) &= \mathcal{N}(0, K_L) & (1)\\ &= (2\pi)^{-\frac{\ell}{2}}|K_L|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}f^T K_L^{-1} f\right) \end{aligned}$$

where K_L is an $\ell \times \ell$ covariance matrix. In particular, the element of K_L is built by means of a covariance function (kernel) $k(x, x')$. A simple example is the standard Gaussian covariance defined as:

$$k(x, x') = c \cdot \exp\left(-\frac{1}{2}\sum_{j=1}^{d} b_j \left(x^j - x'^j\right)^2\right), \theta = \{c, b\} \quad (2)$$

where $b = \{b_j\}_{j=1}^d$ plays the role of characteristic length-scales. c is the kernel over scale. The parameters b and c are initially unknown and are added to a parameter set θ, which is defined as containing all such hyper-parameters.

For a GP model, the marginal likelihood is equal to the integral over the product of likelihood $p(y|f) = \mathcal{N}(f, \sigma^2 I)$ and the prior $p(f|X_L)$, given as:

$$p(y|X_L) = \int p(y|f)\ p(f|X_L)\ df = \mathcal{N}(0, K_L + \sigma^2 I) \quad (3)$$

which is typically thought as the training model of GP. Given some observations and a covariance function, we want to find out the most appropriate θ and σ, and make a prediction on the test data. There are various methods for determining the parameters. A general one is the gradient ascent, which seeks the optimal parameters by maximizing the marginal likelihood.

Given the observations and optimal θ and σ, the prediction distribution of the target value f_* for a test input x_* can be expressed as [18]:

$$p\left(f_* \mid x_*, X_L, y\right) = \mathcal{N}(m_*, c_*) \tag{4}$$

where the predictive mean and variance are:

$$\begin{aligned} m_* &= k_*^T \left(K_L + \sigma^2 I\right)^{-1} y \\ c_* &= k_{**} - k_*^T \left(K_L + \sigma^2 I\right)^{-1} k_* \end{aligned} \tag{5}$$

where k_* is a matrix of covariances between the training data and test data. The matrix k_{**} consists of the covariance of the test data.

IV. Semi-supervised Gaussian Process Regression

As we can see in standard GPr, neither the prior of latent function f (Eq.(1)) nor the predictive distribution (Eq.(4)) contains any information of the unlabeled data. Evidently, to train a accurate GP model, we need to get sufficient training data (labeled data). However, the training data is often difficult and expensive to obtain, while the unlabeled data is relatively easy to collect. Therefore, it appears necessary to modify the standard GP model to make it capable of learning from unlabeled data, and thereby improve the performance of prediction. In this section we present how to effectively use the information of unlabeled data to extend the standard GP model into the semi-supervised framework.

According to semi-supervised smoothness assumption, if two points are close, then so should be the corresponding outputs. Based on this assumption, the unlabeled data should be helpful in regression problem. They can help explore the nearness or similarity between outputs. And the output should vary smoothly with this distance. So, to utilize the unlabeled data, we consider building an adjacent graph to define the nearness between labeled and unlabeled data. Then we attempt to incorporate the graph information into the standard GP probabilistic framework to generate a new probability model for semi-supervised GPr.

A. Prior Condition On Graph

In order to take advantage of the information of unlabeled data, we build an adjacent graph $\mathcal{G} = (V, E)$ on all observed data points $X_D = \{X_L, X_U\}$, to find the adjacent relationship between labeled and unlabeled data, where V is the set of nodes composed by all data points, E is the set of edges between nodes. The graph can be represented by a weight matrix $W = \{w_{ij}\}_{i,j=1}^N$, where $w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\eta}\right)$ is the edge weight between nodes i and j, with $w_{ij} = 0$ if there is no edge.

From the previous section, we can see that regression by GP is a probabilistic approach. Probabilistic approaches to regression attempt to model $p(y|X_D)$. In this case, in order to make the unlabeled data affect our predictions, we must make some assumptions on the underlying distribution of input data. In our work, we attempt to combine the graph information with the GP. Thus, we focus on incorporating a prior of $p(\mathcal{G}|f)$ with the prior of $p(f|X_D)$ to infer a posterior distribution of f condition on the graph $\mathcal{G}$.

Here, we consider the graph $\mathcal{G}$ itself as a random variable. There are many ways to define an appropriate likelihood of the variable $\mathcal{G}$. [21] provides a simple likelihood of observing the graph:

$$p\left(\mathcal{G}|f\right) \propto \exp\left(-\frac{1}{2} f^T \Delta f\right) \tag{6}$$

where Δ is a graph regularization matrix, which is defined as the graph Laplacian here. We can derive Δ in the following way: let $\Delta = \lambda L^{\upsilon}$, where λ is a weighting factor, υ is an integer, and L denotes the combinatorial Laplacian of the graph $\mathcal{G}$. Let $D_{ii} = \sum_j w_{ij}$, the combinatorial Laplacian is defined as $L = D - W$.

Combining the Gaussian process prior $p(f|X_D)$ with the likelihood function Eq.(6), we can obtain the posterior distribution of f on the graph $\mathcal{G}$ as follows:

$$p(f|X_D, \mathcal{G}) = \frac{1}{p(\mathcal{G})} p(\mathcal{G}|f) p(f|X_D) \tag{7}$$

Observably, the posterior distribution Eq.(7) is proportional to $p(\mathcal{G}|f)p(f|X_D)$, which is a multivariate Gaussian as follows:

$$p(f|X_D, \mathcal{G}) = \mathcal{N}\left(0, (K_{DD}^{-1} + \Delta)^{-1}\right) \tag{8}$$

The posterior distribution Eq.(8) will be used as the prior distribution for the following derivation. To proceed further, we have to derive the posterior of f_X independent of graph $\mathcal{G}$. Here X denotes the more general dataset, which contains observed dataset X_D and a set of unseen test data X_T, *i.e.*, $X = \{X_D, X_T\}$. In standard GP, the joint Gaussian prior distribution of f_X can be expressed as follows:

$$p\left(f_X|X\right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{DD} & K_{DT} \\ K_{DT}^T & K_{TT} \end{bmatrix}\right) \tag{9}$$

Then the same as above, the posterior distribution of f_X conditioned on $\mathcal{G}$ is proportional to $p(\mathcal{G}|f_X)p(f_X|X)$, and it is explicitly given by a modified covariance function defined in the following:

$$p\left(f_X|X, \mathcal{G}\right) = \mathcal{N}\left(0, \tilde{K}_{XX}\right) \tag{10}$$

where

$$\tilde{K}_{XX}^{-1} = \begin{bmatrix} K_{DD} & K_{DT} \\ K_{DT}^T & K_{TT} \end{bmatrix}^{-1} + \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \tag{11}$$

Eq.(10) gives a general description that for any finite collection of data X, the latent random variable f_X conditioned on graph $\mathcal{G}$ has a multivariate normal distribution $\mathcal{N}(0, \tilde{K}_X)$, where $\tilde{K}_X$ is the covariance matrix, whose elements are given by evaluating the following kernel function:

$$\tilde{k}\left(x, z\right) = k\left(x, z\right) - k_x^T \left(I + \Delta K\right)^{-1} \Delta k_z \tag{12}$$

in this equation, K is a $N \times N$ matrix of $k\left(\cdot, \cdot\right)$, and k_x and k_z denote the column vector $\left(k\left(x_1, x\right), \ldots, k\left(x_{l+u}, x\right)\right)^T$.

We notice that by incorporating the graph information Δ with the standard GP prior $p(f|X)$, we infer a new prior condition on the graph $\mathcal{G}$ and a graph-based covariance function $\tilde{k}$. In fact this semi-supervised kernel (covariance function) was first proposed by [22] from the Reproducing Kernel Hilbert Space view, and is used for the semi-supervised classification task. In our work, we mainly focus on how to utilize the new prior and the graph-based covariance function to derive the training and predicting distributions for semi-supervised GPr.

B. Objective Functions

Our objective training function for semi-supervised GPr is the marginal likelihood $p(y|X_D, \mathcal{G})$, which is the integral of the likelihood times the prior:

$$p\left(y|X_D, \mathcal{G}\right) = \int p(y|f)\ p(f|X_D, \mathcal{G})\ df \tag{13}$$

Similar with standard GP, the term marginal likelihood refers to the marginalization over the latent function value f. But the difference is that the prior of semi-supervised GP is the posterior obtained by conditioning the original GP with respect to graph $\mathcal{G}$.

According to Eq. 8 and the likelihood $p(y|f) = \mathcal{N}(f, \sigma^2 I)$, the marginal likelihood of the observed target values y is:

$$p\left(y|X_D, \mathcal{G}\right) = \mathcal{N}(0, \Sigma) \tag{14}$$

where $\Sigma = \left(K_{DD}^{-1} + \Delta\right)^{-1} + \sigma^2 I$. This formula can be seen as the training model of our proposed method. We can select the appropriate values of hyper-parameters $\Theta = \{\theta, \sigma\}$ by maximizing the log marginal likelihood $\log p(y|X_D, \mathcal{G})$. The goal is to solve $\hat{\Theta}$ =arg max $\log p(y|X_D, \mathcal{G})$. In learning process we seek the partial derivatives of the marginal likelihood, and use them for the gradient ascent to maximize the marginal likelihood with respect to all hyper-parameters.

After learning the model parameters, we are now confronted with the prediction problem. In the prediction process, given a test data x_*, we are going to infer f_* based on the observed vector y. According to the prior Eq.(10) and Eq.(14), the joint distribution of the training output y and the test output f_* is

$$\left[\begin{array}{c} y \\ f_* \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \Sigma & \tilde{k}_* \\ \tilde{k}_*^T & \tilde{k}_{**} \end{array}\right]\right) \tag{15}$$

Then we can use this joint probability and Eq.(14) to compute the Gaussian conditional distribution over f_*:

$$\begin{aligned} & p\left(f_*|x_*, X_D, y, \mathcal{G}\right) \\ & \quad \propto \exp\left(-\frac{1}{2}\left[y, f_*\right]\left[\begin{array}{cc} \Sigma & \tilde{k}_* \\ \tilde{k}_*^T & \tilde{k}_{**} \end{array}\right]^{-1}\left[\begin{array}{c} y \\ f_* \end{array}\right]\right) \end{aligned} \tag{16}$$

By using the partitioned inverse equations, we can derive the Gaussian conditional distribution of f_* at x_*:

$$p\left(f_*|x_*, X_D, y, \mathcal{G}\right) = \mathcal{N}\left(\hat{\mu}, C\right) \tag{17}$$

where

$$\begin{aligned} \hat{\mu} &= \tilde{k}_*^T \Sigma^{-1} y \\ C &= \tilde{k}_{**} - \tilde{k}_*^T \Sigma^{-1} \tilde{k}_* \end{aligned} \tag{18}$$

This is the key predictive distribution for our proposed semi-supervised GPr method. $\hat{\mu}$ is the mean prediction at the new point and C is the standard deviation of the prediction. For fixed data and fixed hyper-parameters of the covariance function, we can predict the test data from the labeled data and a large amount of unlabeled data.

Note that the graph $\mathcal{G}$ contains the adjacent information of labeled and unlabeled data, and it is helpful for regression according to the smoothness assumption of supervised learning. Then, the knowledge on $p(\mathcal{G}|f)$ that we gain through the unlabeled data carries information that is useful in the inference of $p(y|X_D, \mathcal{G})$ and $p(f_*|x_*, X_D, y, \mathcal{G})$, which is the training probability and predictive distribution for semi-supervised GP regression. Thus, our semi-supervised GPr method can be expected to yield an improvement over supervised one.

V. Regression with Feedback

In the semi-supervised regression, we learn a predictive model from labeled and unlabeled data. Then the output of the unlabeled data can be predicted through the model. In this process, predictive output can be viewed as a kind of experience. Such experience provides the possibility of improving the performance of semi-supervised GPr. Therefore, in this paper, we describe a feedback algorithm, which can pick up the useful prediction of unlabeled data for feeding back into the labeled dataset and re-train the model iteratively.

In a predictive system, we can not affirm that all the predictions of unlabeled data could be correctly predicted. For this reason, not all the predictions are helpful for re-training and we need to pick up the useful ones from them. Here we call one useful prediction a confident prediction. Now we have a problem that what the confident prediction is. Intuitively, if a labeled example can help to decrease the error of the regressor on the labeled data set, it should be the confident labeled data. Therefore, in each learning iteration of feedback, the confidence of unlabeled data point x_u can be evaluated using a criterion of:

$$E_{x_u} = \sum_{x_i \in X_L} \left(\left(y_i - M(x_i)\right)^2 - \left(y_i - M'(x_i)\right)^2\right) \tag{19}$$

here, M is the original semi-supervised regressor trained by the labeled dataset (X_L, y_L) and unlabeled dataset X_U, while M' is the one re-trained by the new labeled dataset $\{(X_L, y_L) \cup (x_u, \hat{y}_u\}$ and unlabeled dataset $\{X_U - x_{u'}\}$. Here x_u is an unlabeled data point while $\hat{y}_u$ is the real-valued output predicted by the original regressor M, i.e. $\hat{y}_u = M(x_u)$. The first term of Eq.(19) denotes the mean squared error (MSE) of the original semi-supervised regressor on labeled dataset, and the second term is expressed the MSE of the regressor utilizing the information provided by $(x_u, \hat{y}_u)$ on the labeled dataset. Thus, $(x_u, \hat{y}_u)$ associated with the biggest positive E_{x_u} can be regarded as the most confident labeled data. In other words, If the value of E_{x_u} is positive, it means utilizing $(x_u, \hat{y}_u)$ is beneficial. So we can use this unlabeled data paired with its prediction as labeled data in the next round of model training. Otherwise, $(x_u, \hat{y}_u)$ is not helpful to train models, and will be omitted. Then the x_u should remain in the unlabeled dataset X_U .

TABLE I. ALGORITHM OF FEEDBACK

```
Input: Labeled dataset($(X_L, y_L)$), Unlabeled dataset $X_U$,
       Learning iterations $T$, Initial parameters set $InitPara$
Output: Prediction model $M$
Step1: Training model
$M \leftarrow Semitrain(X_L, y_L, X_U, InitPara)$
Step2: Choosing and feedback
for $t = 1 : T$ do
  Create pool $X_{U'}$ by randomly selecting data points from $X_U$
  for each $x_u \in X_{U'}$ do
    $\hat{y_u} \leftarrow M(x_u)$
    $M' \leftarrow Semitrain\left((X_L, y_L) \cup (x_u, \hat{y_u}), \{X_U - x_u\}, InitPara\right)$
    $E_{x_u} \leftarrow \sum_{x_i \in X_L} \left((y_i - M(x_i))^2 - (y_i - M'(x_i))^2\right)$
  end for
  for each $E_{x_u} > 0$ do
    $(X_L, y_L) \leftarrow (X_L, y_L) \cup (x_u, \hat{y_u})$
    $X_U \leftarrow \{X_U - x_u\}$
    $M \leftarrow M'$
  end for
  $M \leftarrow Semitrain(X_L, y_L, X_U, InitPara)$
end
```

The pseudo code of our feedback framework is shown in Table I, where the function $Semitrain$ returns a semi-supervised GP regressor. The learning process stops when the maximum number of learning iterations, T, is reached, or there is no unlabeled data.

VI. EXPERIMENTS

In this section, we firstly evaluate the performance of the proposed semi-supervised GPr (SemiGPr) on some regression datasets, and make a direct comparison to its standard version (GPr). Then we show the experimental results of SemiGPr with the feedback algorithm (named FdGPr). Finally, we introduce the clustering framework, and empirically demonstrate the exclusion time and accuracy of the local SemiGPr extension by this framework.

There are $d + 4$ hyper-parameters in SemiGPr: kernel length-scales $b = \{b_i\}_{i=1}^d$, where d is the dimension of input x, kernel over scale c, noise σ and edge weight length-scale η. In our experiment, we select the appropriate values of $\{b, c, \sigma\}$ by maximizing the marginal likelihood. To reduce the computing complexity, we fix $\eta = 10$ for all datasets. 4-fold cross validation is performed on each dataset and all the results are averaged over 40 runs of the algorithm.

The datasets used to evaluate the performance of our method are summarized in Table II. The examples contained in the artificial dataset Friedman is generated from the function: $y = \tan^{-1}\left(x_2 x_3 - 1/x_2 x_4\right)/x_1$. The constraint on the attribute is: $x_1 \sim U[0, 100]$, $x_2 \sim U[40\pi, 560\pi]$, $x_3 \sim U[0, 1]$, $x_4 \sim U[1, 11]$. Gaussian noise term is added to the function. The real-world datasets are from the UCI machine learning repository and StatLib.

In our experiment, for each dataset, we randomly choose 25% of the examples as test data, while the remaining are training data. We take 10% of the training data as labeled examples, and the remaining is used as the set of unlabeled examples. Note that all the datasets are normalized to the range $[0, 1]$.

A. Algorithmic Convergency

In this paper, we estimate hyper-parameters by using the gradient descent method to minimize the following log marginal likelihood.

$$-\log p\left(y|X, \mathcal{G}\right) = \frac{1}{2} y^T \Sigma^{-1} y + \frac{1}{2} \log|\Sigma| + \frac{N}{2} \log 2\pi \quad (20)$$

TABLE II. DATASETS USED FOR SEMIGPR. D IS THE FEATURE; N DENOTES THE SIZE OF THE DATA.

Dataset	Friedman	wine	chscase	no2
D	4	11	6	7
N	3000	1599	400	500
Source	Artificial	UCI	Statlib	Statlib
Dataset	kin8nm	triazines	pyrim	bodyfat
D	8	60	27	14
N	2000	186	74	252
Source	UCI	UCI	UCI	Statlib

Firstly, we discuss the convergence of the above training objective function. In Figure 1, we show how the objective function value decreases as a function of the iterations on triazines (left) dataset and no2 (right) dataset. The result of triazines shows a typical convergence process. As the number of iterations is increasing, the objective function value is decreasing smoothly. Meanwhile, the objective function value of no2 is converged in two stages. From the results, we can see that the objective function value decreases with the increase of the number of iterations and the iterative procedure guarantees a local optimum solution for the objective function in Eq.(20). According to our offline experiments, generally, the objective function converges after about 30-40 iterations for the datasets in Table II.

B. Efficiency of Unlabeled Data

To verify the SemiGPr model can take advantage of unlabeled data, for a fixed number of labeled data, we vary the number of unlabeled examples, and plot the mean squared error (MSE) for dataset triazines and no2. The corresponding curves are shown in Figure 2, where the dotted line and solid line indicate the predictive errors on unlabeled dataset and test dataset respectively. Note that when the proportion of unlabeled data is 0%, the result denotes the MSE of standard GPr. The figure shows that the proposed semiGPr algorithm has lower MSE compared to the standard GPr both on unlabeled and test dataset. Moreover, as the proportion of unlabeled examples increases, the advantage of semiGPr increases further. From this result, we can conclude that SemiGPr may bring extra advantage by utilizing the unlabeled data for model training. In other words, the unlabeled data provides some useful information, and our semi-supervised algorithm can make use of this information to improve the predictive accuracy.

While we observe a significant performance improvement of the proposed algorithm by using unlabeled examples, the unlabeled examples are not always helpful. For example, for data no2 (right figure of Figure 2), when the proportion of unlabeled data goes from 30% to 50%, the error rates are increased instead of reduced. The same happens to triazines when the size of unlabeled dataset goes from 90% to 100%. It is of interest to find out the cause of the negative effect of the unlabeled data experimentally in the future.

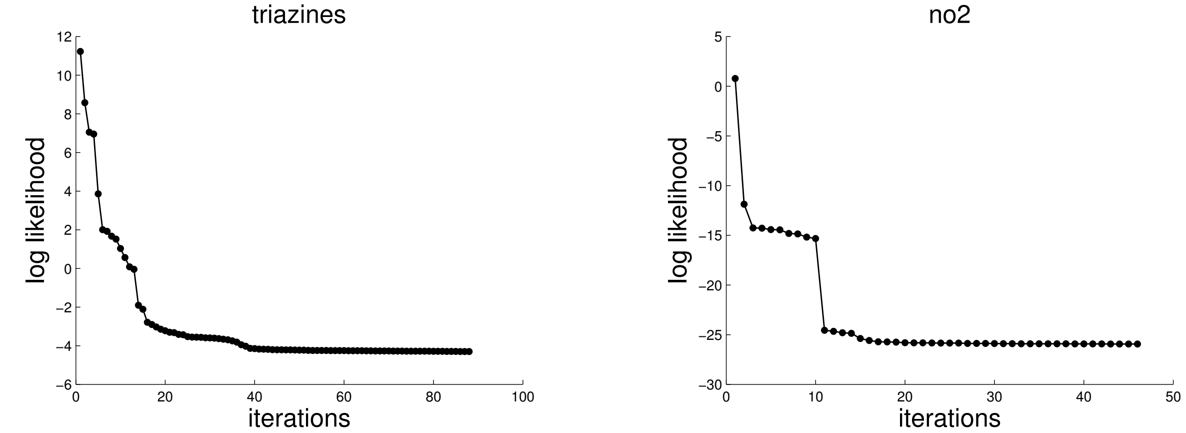


Fig. 1. log likelihood decreases along with the increase of the iteration No. for the triazines (left) and the no2 (right).

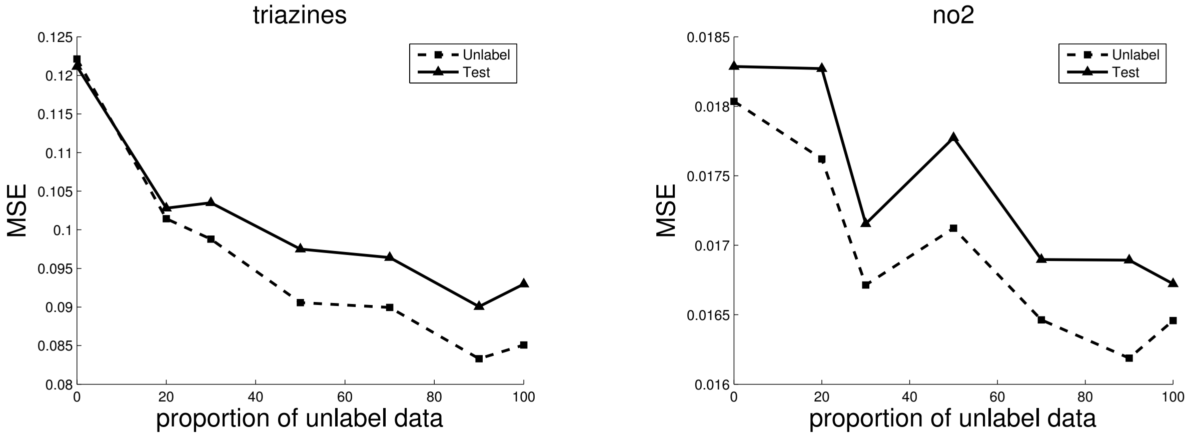


Fig. 2. Performance of SemiGPr as a function of number of unlabeled examples.

C. Evaluation of Regression Accuracy

To further clarify the effect of the proposed method, we compare the MSE between SemiGPr and GPr. The comparative results are summarized in Table III. The above value is the performance on unlabeled dataset, and the following value is on test dataset. In this experiment, we consider GPr as the baseline and compare the performance of SemiGPr with it. The improvements are also listed in the table. In addition to the average MSE, we test the significance of the performance difference between SemiGPr and GPr using a paired t-test on the MSE values. The differences are significant with a paired t-test at the 0.05 level, and the results with significant improvement in the table are bold-faced.

The result in Table III shows that our method SemiGPr performs as well as or better than the standard GPr in terms of the regression accuracy. We can observe that SemiGPr leads to improvements in most of the datasets, and the differences are significant in about half of the datasets. From the comparison, we can conclude that using the unlabeled training data with our semi-supervised regression framework, the GP regression accuracy can be improved. On some of the datasets like chscase, the precision of SemiGPr did not have a significant improvement over the standard one. There are two possible reasons for this result. One is the poor hyper-parameter choices made in optimization process. The other one is the negative effect of the unlabeled data as shown in the previous experiment.

D. Comparison with other methods

To further evaluate the performance of SemiGPr, we compare our results with other semi-supervised regression methods. In the first experiment, the co-training method (COREG) presented in [?] is compared. The code and documentation of COREG are available at $http : //lamda.nju.edu.cn/code/\_COREG.ashx$. All the experimental setting of COREG is the same as that of SemiGPr, i.e., the same splitting of the training and testing sets and preprocessing methods, and 40 randomly runs of the algorithm for each dataset. The obtained results are summarized in Table IV. We perform a paired t-test at the 5% significance level, and the results with obvious improvement in the table are bold-faced. In general, we observe that our method achieves a smaller error on all of the datasets compared to COREG. In particular, on Wine and kin8nm datasets, we observe a significant performance improvement of the MSE over COREG. It confirms the conclusion that our semi-supervised method can take advantage of the unlabeled data and it is effective even when only a limited amount of labeled data is available.

In the second experiment, in order to illustrate the difference between our method and the one proposed in [26],

TABLE III. Comparison of SemiGPr with the standard GPr on different datasets.

Dataset	Friedman	Wine	chscase	no2	kin8nm	bodyfat	pyrim	triazines
GPr	0.0113	0.0196	0.0273	0.0180	0.0136	0.0026	0.0524	0.1215
	0.0114	0.0205	0.0268	0.0183	0.0134	0.0061	0.0544	0.1205
SemiGPr	0.0101	0.0190	0.0264	0.0161	0.0131	0.0026	0.0359	0.0843
	0.0102	0.0199	0.0265	0.0164	0.0132	0.0027	0.0495	0.0925
Improv.	10.62%	**3.06%**	3.30%	**10.56%**	**3.68%**	0%	**31.49%**	**30.62%**
	10.53%	**2.93%**	1.12%	**10.38%**	1.49%	55.74%	9.01%	**23.24%**

TABLE IV. Comparison of SemiGPr with co-training methods.

Dataset	Friedman	Wine	chscase	no2	kin8nm
SemiGPr	0.0102	**0.0199**	0.0265	0.0164	**0.0132**
COREG	0.0115	0.0214	0.0282	0.0166	0.0190

TABLE V. Comparison of SemiGPr with SSMTR (SSRT:supervised single-task regression which uses one GP for each task).

Method	SSTR	SSMTR	SemiGPr
robot arm	1.0228 ± 0.1318	0.3810 ± 0.1080	0.8389
	1.0270 ± 0.1450	0.3905 ± 0.1123	0.8710
school	1.2914 ± 0.3146	1.0506 ± 0.2804	1.3266
	1.3240 ± 0.3274	1.0612 ± 0.2813	1.0723

we run experiments on exactly the same datasets of [26], following precisely their preprocessing and testing methods, where in Robot arm dataset, 2000 data points are selected independently for each task, with 1% as labeled data, 10% as unlabeled data and the remaining as test data, and in School dataset, for each task, 2% of the data is selected as labeled data, 20% as the unlabeled data and the rest as test data. Here, we use the systematic sampling method as the selection method. The normalized mean squared error, which is defined as the mean squared error divided by the variance of the test output, is calculated as the performance measure. The results are averaged over 10 runs of the algorithm. In Table V we report the test normalized mean squared error for two multi-task regression datasets.

It turns out that Zhang's method of SSMTR uses a very similar idea: constructing an adjacency graph and incorporating the prior of the graph with the GP prior to generate a semi-supervised data-dependent kernel function. Actually we derived the marginal likelihood and predictive distribution of the GP from different routes. As we discussed earlier, the major difference between two methods is that we have totally different training and prediction models.

From the results of this experiment in Table V we can obtain the merits and demerits of these two different models. The result shows that SSMTR achieves a smaller error on robot arm dataset compared to our method because Zhang's method considers the relevance among tasks. Their model consists of GP and a common prior on the parameters for all tasks, and the common prior can model the relevance well. The robot arm dataset contains 7 tasks which are 7 joint toques of the robot arm. The 7 joint toques have strong association with each other, which means the 7 tasks have high relevance. Therefore, for such a multi-task regression, it's better to learn a multi-task model rather than building a single-task model for each task independently. However, for the second dataset (School score), although the tasks have some relevance with each other, our method still performs as well as the multi-task method SSMTR. In this dataset, the examination scores of students between different schools are related with the difficulty of the examination. In Zhang's paper, to model this latent relevance, they impose a common Gaussian prior $\theta^i \sim \mathcal{N}(m_\theta, \Sigma_\theta)$ on the kernel parameters for all tasks. On the other hand, our method is proposed for single task and can not model this relevance well. However, the common prior has no effect on a single task. Because when there is only one task in a dataset, the common prior becomes a fixed value. Therefore, it may be interesting in the future to compare which performs better for single-task.

Another major difference between two methods lies in hyper-parameter optimization. In Zhang's method, the number of parameters to estimate is large, since all the tasks are modeled in one formulation, and the number of parameters increases with the tasks. Because of the large number of parameters, it is difficult to estimate the optimal values simultaneously. So the parameters are optimized through an alternating optimization algorithm. And this could cause a computational problem for large multi-tasks datasets. However, in our work, the parameters of each task are estimated separately by maximizing the log-likelihood. Therefore, our work can be parallelized easily for a multi-task. It will also be interesting in the future to compare which performs better for hyper-parameter optimization and which saves training time.

E. Results of Feedback Algorithm

In this part, two of the datasets used in SemiGPr are presented to demonstrate the effectiveness of the SemiGPr extended by the feedback algorithm, which is denoted by FdGPr. Experimental setting is the same as the previous subsection.

To clarify unlabeled examples and their predictions really contain some valuable information and our feedback algorithm can utilize such information to improve the predictive accuracy, we plot the MSE of FdGPr for different iteration numbers. The results are shown in Figure 3. The dot line denotes the MSE on the unlabeled dataset, and the solid line is the result of the test dataset. The left figure is the result of dataset no2 and the right one is that for chscase. Note that when the feedback iteration is 0, the result denotes the MSE of SemiGPr.

From the figures we can see that when the iteration number is increased, the feedback algorithm cuts the error

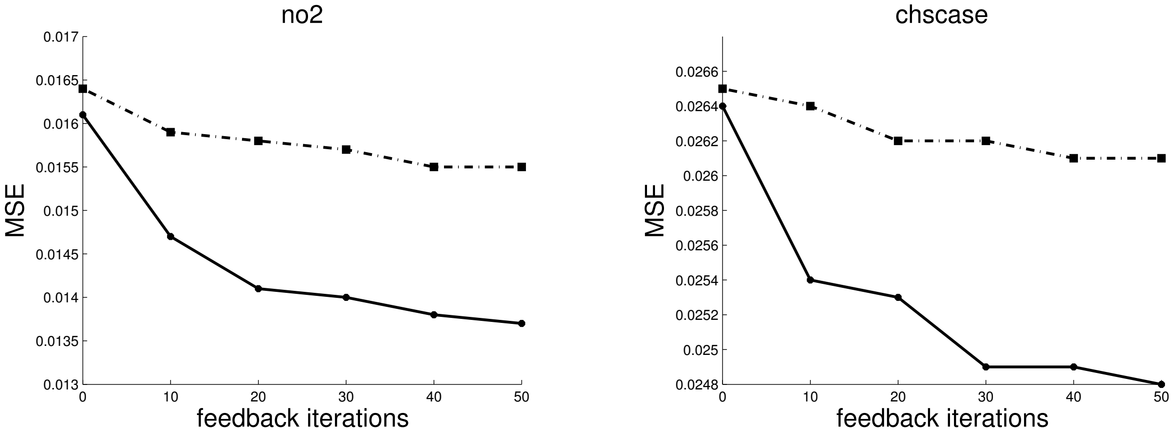


Fig. 3. The effect of different feedback iterations on unlabeled and test dataset.

rate drastically over SemiGPr. The results show clearly that the unlabeled examples and their predictions have a beneficial effect on model learning. From the experiment, larger iteration number almost always produces better results, while considering the computational cost, the iteration T should be set to 20. Although FdGPr achieves a comparable performance to a non-feedback baseline on the unlabeled dataset, it does not have a significant improvement over the other ones on the test dataset. Therefore we should point out that the feedback algorithm makes our work transductive, and we should find a new metric to select predictions of unlabeled examples to improve the performance on test dataset in future work. From this result, we can make a conclusion that by utilizing feedback information, FdGPr makes performance improvements over the other methods, especially on unlabeled data.

F. Extension by Clustering Regression Framework

A significant problem with GP is that it's computationally expensive to carry out the necessary matrix computation ($O(N3)$). To address this problem, we give a further extension of the SemiGPr by a clustering regression framework and discuss the possibility of improving the efficiency while keeping the accuracy.

Indeed, some regression methods with similar idea have already been proposed. For example, [25] proposed a regression clustering algorithm to solve the complex distribution regression problem. The proposed algorithm updates the data in each cluster by using a regression error. Then the clusters and the corresponding regression functions can be obtained simultaneously. This method is effective for the dataset that has multiple tasks within it. In addition, in [13] and [8], excellent results have been obtained on some specific datasets by clustering the input data into several parts, and learning a regression model inside each cluster. In these studies, the accuracy of combining the clustering and regression has been discussed for the specific dataset, such as mixture distribution dataset and multiple spatial dataset, while we focus on empirically demonstrating the execution time and accuracy on general regression dataset. Besides, in above studies supervised approaches have been used for regression, but in our work we take advantage of the proposed semi-supervised regression.

TABLE VI. ALGORITHM OF FUZZY C-MEANS CLUSTERING

```
**Input:** $X = \{X_L, X_U\}$, $P_{init}$, $Params(m, \epsilon, A)$
**Output:** P, C
**Repeat**
**Step1:** Compute the centroid of cluster
**for** each cluster $j$, $1 \leq j \leq K$ **do**
    $c_j^{(t)} \leftarrow \frac{\sum_{i=1}^{N} (p_{ij}^{(t-1)})^m x_i}{\sum_{i=1}^{N} (p_{ij}^{(t-1)})^m}$ (a)
**Step2:** Calculate the distances from data to center
**for** each data point $x_i$, $1 \leq i \leq N$ **do**
    $D_{ij}^2 \leftarrow (x_i - c_j)^T A (x_i - c_j), 1 \leq j \leq K$ (b)
**Step3:** Update the partition matrix
    $p_{ij}^{(t)} \leftarrow \frac{1}{\sum_{k=1}^{K} (D_{ij}/D_{ik})^{2/(m-1)}}$ (c)

**Until** $\parallel P^{(t)} - P^{(t-1)} \parallel < \epsilon$
```

Our clustering regression framework consists of three stages: 1) data partition, 2) training models and 3) output prediction. The pseudo code for the different phases of this framework is shown in Table VII.. The first step is to partition the input data into several clusters. General clustering methods, for example k-means, divide the data into distinct clusters, where each data point belongs to exactly one cluster. However, this constraint is prone to cause unsuitable clustering results in the boundary areas among different clusters. Consequently, in this paper data partition is performed using a soft clustering model named fuzzy c-means clustering (FCM) [2].

The FCM algorithm attempts to partition a finite collection of N elements $X = \{X_L, X_U\} = \{x_1, \ldots, x_N\}$ into a collection of K fuzzy clusters with respect to some given criterion. Here, the X_L and X_U denote the labeled input set and the unlabeled input set separately. Given a finite set of data X, the algorithm returns a list of K cluster centers $C = \{c_1, \ldots, c_K\}$ and a $N \times K$ partition matrix $P = p_{ij} \in [0, 1]$, $i = 1, \ldots, N$, $j = 1, \ldots, K$, where each element p_{ij} can be interpreted as the probability that the element x_i belongs to cluster j.

Table VI provides the learning process of FCM. Given X and the initial partition matrix P, the FCM algorithm first computes the cluster centroid c_i for each cluster, using the formula Eq.(a) in Table VI. The centroid of a cluster is the mean of all points, weighted by their degree of belonging to the

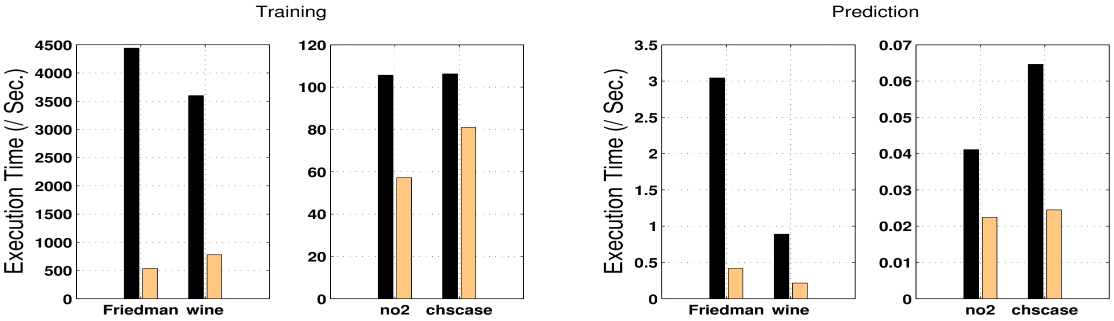


Fig. 4. Comparison of execution time

TABLE VII. PSEUDO CODE OF CLUSTERING FRAMEWORK

```
Input: $X, y_L, x_t, P_{init}, Params, hypara$
Output: P, C, M
Step1: Data Partition
$[P, C] = FuzzyCmeans(X, P_{init}, Params)$
Step2: Training Model
for each cluster $j (j = 1 : K)$ do
    $M_j \leftarrow Semitrain(X^j, y_L^j, hypara)$
end for
Step3: Output Prediction
Compute the partition coefficient $p_t$ of $x_t$
$w_{tj} \leftarrow p_{tj} / \sum_{p_t > \delta} p_t$
$\hat{y}_t \leftarrow \sum_{x_t \in j} w_{tj} * M_j(x_t)$
```

cluster. Then it calculates the distances D_{ij} from data point x_i to the cluster center c_j. Finally, for each data point, it updates the coefficients of being in the clusters. It is repeated until the convergence condition is satisfied.

Through the algorithm presented in Table VI, we can get the partition matrix P, where the rows indicate the probabilities that the data point x belongs to each cluster, and the columns denote the probabilities that all the data points X are partitioned into cluster j. But the final goal of clustering is to calculate an indicator matrix $Z = z_{ij} \in \{0, 1\}$, $i = 1, \dots, N$, $j = 1, \dots, K$. Here, the z_{ij} is one if x_i is assigned to the corresponding cluster j, or zero otherwise. The general idea of obtaining Z from P is that setting the maximum probability of each row to be 1, and the rest are 0. But it becomes the same as hard clustering to some extent. To gain fuzzy clusters, in this paper we exploit a threshold δ to filter the partition matrix. For example, set δ to be 0.4, then the corresponding z_{ij} is one if $p_{ij} > 0.4$, and zero otherwise. Thus, for a data point, it may not be assigned to only one cluster but to the clusters that have probability bigger than δ. Through adjusting the threshold δ, we can control how much clusters may overlap.

Several clusters with overlaps are obtained by the data partition step. Then in training step, local semi-supervised GPr model M (Eq.14) is trained for each cluster. Finally in predicting step, the prediction of a given data point x_t is a weighted combination of the predictions of the individual local models given by $\hat{y}(x_t) = \sum_{j=1}^{K} P_{x_t}(j) * M_j(x_t)$, where $P_{x_t}(j)$ denotes the weight of the model j. It effectively equals to the partition matrix that tells the probability of element x_t belonging to cluster j. And $M_j(x_t)$ represents the prediction of input x_t by using the model M_j, the formula of which is Eq. 18.

TABLE VIII. RESULTS OF CLUSTERING REGRESSION FRAMEWORK.

Dataset	Friedman	Wine	chscase	no2
SemiGPr	0.0101	0.0190	0.0264	0.0161
	0.0102	0.0199	0.0265	0.0164
CSGPr	0.0106	0.0220	0.0256	0.0154
	0.0107	0.0232	0.0261	0.0160

To evaluate the performance of the clustering regression framework (named FCMGPr), we experiment on the datasets described in Table II. The parameters chosen for the FCM algorithm remain unchanged for each dataset, $m = 2$, $\epsilon = 10^{-6}$, and A is a diagonal matrix with size $d \times d$, where d is the dimension of data X. Here, we set the threshold δ to be 0.4. And if the data points in a dataset are more than one thousand, then the number of clusters K equals to 4, if not, $K = 2$. The experimental setting of regression part is the same as the evaluation of SemiGPr.

Firstly, we compared the accuracy between SemiGPr and FCMGPr. The MSE results on the unlabeled and test data are shown in Table VIII. From the results we can see that the FCMGPr performs better than the semi-supervised one on no2 and chscase datasets. One of the reasons is that the local models are more flexible than a global one. In other words, constructing model locally can capture the details of data better than applying a global model across entire dataset. In addition, making predictions by weighted combination can help to avoid the inaccurate results due to incorrect clustering. However, there is a problem with clustering regression framework: if the amount of labeled data is too small in a cluster, particularly high-dimensional data, it will be easy to make poor hyper-parameter choices or occur under-fitting. Then it results in bad accuracy for FCMGPr. This is an explanation of the results on wine and Friedman datasets, where the local model did not have an improvement over the semi-supervised one.

Secondly, we tested the exclusion time of SemiGPr and FCMGPr. The results are shown in Figure 4, where the left two figures are the comparison of training time and the right two figures show the results of predicting time. As we can see in Figure 4, for all four datasets, both the training time and the predicting time of FCMGPr are reduced over SemiGPr. From these results we can conclude that by using the clustering regression framework, the computational efficiency of the semi-supervised GPr can be greatly improved and it also has the possibility for improving the prediction accuracy.

VII. Conclusion

In this paper we presented and evaluated a semi-supervised GPr by incorporating an adjacent graph within the standard GP probabilistic framework. Through exploring the standard GP to semi-supervised setting, we can learn a regression model from only small number of expensive labeled data and a large amount easily obtained unlabeled data. Moreover, we presented a feedback algorithm, which can choose the confident prediction for feedback to further improve the performance. Furthermore, to solve the computational problem of GP, we also gave a further extension of the semi-supervised GPr by a clustering regression framework. The experimental results indicate that our semi-supervised regression approach can improve the prediction accuracy. Besides, by choosing the confident prediction for feedback, it brings a significant improvement in the prediction accuracy over a non-feedback baseline. The extension by the clustering regression framework is successful in reducing the exclusion time.

In the experiments, we compared SemiGPr with some state-of-art methods. There also exist some other semi-supervised regression methods, such as regularization regression method [14], propagable graph method [16]. However, because of the different experimental settings, we could not compare the proposed method with them. Future work should include implementing these methods and empirical comparisons with them. We will also apply our scheme to harder regression tasks. Although the results of feedback extension were encouraged, it is noted that the algorithm has high time complexity due to the re-training of SemiGPr. Therefore, in the future work a new feedback criterion would need to be explored in order to obtain more accurate predictions but spending less computational cost.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Balasko, B., Abonyi, J., and Feil, B., *Fuzzy clustering and data analysis toolbox*, Department of Process Engineering, University of Veszprem, Veszprem, Computer software manual, 2005.

[3] Belkin, M., Niyogi, P., and Sindhwani, V., *Manifold regularization: A geometric framework for learning from labeled and unlabeled examples*, Journal of Machine Learning Research, 7, 2399–2434, 2006.

[4] Blum, A., and Mitchell, T., *Combining labeled and unlabeled data with co-training*, In Proceedings of the 11th Annual Conference on Computational Learning Theory, pp. 92–100, 1998.

[5] Brefeld, U., Gärtner, T., Scheffer, T., and Wrobel, S., *Efficient co-regularised least squares regression*, In Proceedings of the 23rd International Conference on Machine Learning, pp. 137–144, 2006.

[6] Chapelle, O., Schölkopf, B., and Zien, A., *Semi-supervised learning*, MIT Press, 2006.

[7] Chen, T., and Ren, J., *Bagging for gaussian process regression*, Neurocomputing, 72(7-9), pp. 1605–1610, 2009.

[8] Das, K., and Srivastava, A. N., *Block-gp: Scalable gaussian process regression for multimodal data*, In Proceedings of the 10th IEEE International Conference on Data Mining, pp. 791–796, 2010.

[9] Hanneke, S., and Roth, D., *Iterative labeling for semi-supervised learning*, Tech. Rep. No. UIUCDCS-R-2004-2442, Computer Science Department, University of Illinois at Urbana-Champaign, 2004.

[10] Kim, K. I., Steinke, F., and Hein, M., *Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction*, In Proceedings of the 23th Annual Conference on Neural Information Processing Systems, pp. 979–987, 2009.

[11] Lafferty, J., and Wasserman, L., *Statistical analysis of semi-supervised regression*, In Proceedings of the 21th Annual Conference on Neural Information Processing Systems, pp. 801–808, 2007.

[12] Lawrence, N. D., Seeger, M., and Herbrich, R., *Fast sparse gaussian process methods: The informative vector machine*, In Proceedings of the 16th Annual Conference on Neural Information Processing Systems, pp. 609–616, 2002.

[13] Lazarevic, A., Pokrajac, D., and Obradovic, Z., *Distributed clustering and local regression for knowledge discovery in multiple spatial databases*, In Proceedings of the 8th European Symposium on Artificial Neural Networks, pp. 129–134, 2000.

[14] Lin, B., Zhang, C., and Xiaofei, H., *Semi-supervised regression via parallel field regularization*, In Proceedings of the 25th Annual Conference on Neural Information Processing Systems, pp. 433–441, 2011.

[15] Luo, J., Chen, H., and Tang, Y., *Analysis of graph-based semi-supervised regression*, In Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, pp. 111–115, 2008.

[16] Ni, B., Yan, S., and Kassim, A. A., *Learning a propagable graph for semisupervised learning: Classification and regression*, IEEE Transactions on Knowledge and Data Engineering, 24(1), pp. 114–126, 2012.

[17] Rasmussen, C. E., and Ghahramani, Z., *Infinite mixtures of gaussian process experts*, In Proceedings of the 16th Annual Conference on Neural Information Processing Systems, pp. 881–888, 2002.

[18] Rasmussen, C. E., and Williams, C. K. I., *Gaussian processes for machine learning*, MIT Press, 2006.

[19] Rosenberg, C., Hebert, M., and Schneiderman, H., *Semi-supervised self-training of object detection models*, In Proceedings of the 7th IEEE Workshops on Application of Computer Vision, pp. 29–36, 2005.

[20] Seeger, M., Williams, C. K., and Lawrence, N. D., *Fast forward selection to speed up sparse gaussian process regression*, In Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics, 2003.

[21] Sindhwani, V., Chu, W., and Keerthi, S. S., *Semi-supervised gaussian process classifiers*, In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1059–1064, 2007.

[22] Sindhwani, V., Niyogi, P., and Belkin, M., *Beyond the point cloud: from transductive to semi-supervised learning*, In Proceedings of the 22nd International Conference on Machine Learning, pp. 824–831, 2005a.

[23] Sindhwani, V., Niyogi, P., and Belkin, M., *A co-regularization approach to semi-supervised learning with multiple views*, In Proceedings of the ICML Workshop on Learning with Multiple Views, pp. 74–79, 2005b.

[24] Snelson, E., and Ghahramanin, Z., *Sparse gaussian processes using pseudo-inputs*, In Proceedings of the 19th Annual Conference on Neural Information Processing Systems, pp. 1257–1264, 2005.

[25] Zhang, B., *Regression clustering*, In Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 451–458, 2003.

[26] Zhang, Y., and Yeungn, D. Y., *Semi-supervised multi-task regression*, In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 617–631, 2009.

[27] Zhou, Z. H., and Li, M., *Semisupervised regression with cotraining-style algorithms*, IEEE Transactions on Knowledge and Data Engineering, 19(11), pp. 1479–1493, 2007.

[28] Zhu, X., *Semi-supervised learning literature survey*, Tech. Rep. No. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[29] Zhu, X., Ghahramani, Z., and Lafferty, J. D., *Semi-supervised learning using gaussian fields and harmonic functions*, In Proceedings of the 20th International Conference on Machine Learning, pp. 912–919, 2003.

Influence of Nitrogen-di-Oxide, Temperature and Relative Humidity on Surface Ozone Modeling Process Using Multigene Symbolic Regression Genetic Programming

Alaa F. Sheta
Software Engineering Department
Zarqa University
Zarqa, Jordan

Hossam Faris
Business Information Technology Department
The University of Jordan
Amman, Jordan

***Abstract*—Automatic monitoring, data collection, analysis and prediction of environmental changes is essential for all living things. Understanding future climate changes does not only helps in measuring the influence on people life, habits, agricultural and health but also helps in avoiding disasters. Giving the high emission of chemicals on air, scientist discovered the growing depletion in ozone layer. This causes a serious environmental problem. Modeling and observing changes in the Ozone layer have been studied in the past. Understanding the dynamics of the pollutants features that influence Ozone is explored in this article. A short term prediction model for surface Ozone is offered using Multigene Symbolic Regression Genetic Programming (GP). The proposed model customs Nitrogen-di-Oxide, Temperature and Relative Humidity as the main features to predict the Ozone level. Moreover, a comparison between GP and Artificial Neural Network (ANN) in modeling Ozone is presented. The developed results show that GP outperform the ANN.**

Keywords: Air pollution; Surface Ozone; Multigene Symbolic Regression; Genetic Programming; Multilayer perceptron neural network; Prediction.

I. Introduction

Tropospheric ozone is an air pollution which causes serious human health problems. The insufficient adherence to the international standard air quality trends, growth of industrialized activities and the emitting of various types of gasses such as carbon monoxide (CO), nitrogen oxides (NO_x), Sulphur dioxide (SO_2), and Particle Pollution (PM_{10}) and ($PM_{2.5}$) in the air without any concern about the impact on human health became a common problem worldwide. These behaviors cause a rise to the earth temperature and affect many meteorological variables [1], [2].

The role of stratospheric ozone in the air is to filter out the greatest portion of the sun possibly harmful shortwave the ultraviolet (UV) radiation. This means that the depletion of ozone allows more UV emissions to touch the earths surface. Many studies proved that these UV emissions could have severe impacts on human beings, animals and plants [3]. In [4] authors explored the dramatic effects of UV radiation on the eye and the skin. Higher temperatures associated climate change possibly will lead, among numerous other effects, to increasing rate of skin cancer. The influence of ambient ozone on human health was studied for fifty US cities for five summers was presented in [5]. Countries such as New Zealand developed many studies on air quality to estimate the likely health problem which may be encountered and decide where emissions should be condensed to improve air quality. In [6], a published report studied the influence of CO, nitrogen dioxide (NO_2), SO_2, O_3, and benzene and benzo(a)pyrene (BaP) in air.

In the past, researchers proposed different types of models to forecast the concentrations of pollutants. Some of these models are statistical based like Autoregressive-moving-average (ARMA) models and linear regression models [7]–[10]. Recently, a more attention was given to machine learning techniques based models such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) [2], [11]–[15] for developing forecasting models.

In this work, Multigene Symbolic Regression GP is used to develop short term prediction model of surface Ozone. The proposed model can predict the mean surface Ozone based on limited number of attributes. They are the Nitrogen-di-oxide, temperature and relative humidity. The Multigene GP has some advantages over other techniques like ANNs such as; producing compact mathematical models that have explanation power and easy to evaluate. A complete comparison between both techniques on solving the modeling problem is presented.

This paper is organized as follows. An overview of the ANN technique is presented in Section II. GP as an evolutionary computation technique is presented in Section III. The evaluation criterion adopted to check the performance of the developed models are presented in Section IV. The area of study considered with detailed information about data collection is discussed in V. Section VI provides the experimental setup and results of the two developed models of the Ozone based ANN and Multigene Symbolic Regression GP.

II. Multilayer Perceptron ANN

ANN was first defined as an information-processing system. This system has large number of simple processing units called ”neurons”. These neurons interconnect by sending and receiving signals which activate the neurons connected to it. A huge number of these neurons constitute a neural network. ANN is distinguished by certain performance characteristics such as its architecture, its training algorithm and the activation function. In this work, we investigate the application of multilayer feedforward neural network which is one of the most common types of neural networks applied for function approximation and prediction [13], [16], [17]. In MLP-ANN, neurons are arranged in layers (input, hidden and output layers). The information in feedfoward MLP-ANN flows in only one forward direction, from the input layer, through the hidden layers to the output layer [18]. Figure 1 depicts an example of a feedfoward MLP designed for Ozone prediction. In this example, the MLP has four neurons in a single hidden layer.

A. Learning algorithm

To adjust ANN weights such that the learning process achieved its goal by modeling the relationship between the inputs and output we need a learning algorithm. One of the very famous learning algorithms is the backpropagation (BP) learning algorithms. BP works by adjusting a cost function to minimize the error difference between the actual output and the ANN output. This function could be simply the sum of the error square. The learning process can be split into number of phases as below:

1) **Hidden layer**:
Assume we have a set of input-output measurements in the form of x_i, y_i. The inputs x_i are always presented to the input layer, then pass to the hidden layer weighted by the weights w_{ij}. The hidden layer always have a nonlinear function known as sigmoid function (see Equation 1). The output of each neuron in the hidden layer is the summation function presented in Equation 2.

$$\begin{aligned} y_j &= \phi(S_j) \\ \psi(x) &= \frac{1}{1+e^{-x}} \end{aligned} \tag{1}$$

$$S_i = w_0 + \sum_{i=1}^{n} w_{ij} x_i \tag{2}$$

where $i = 1, 2 \ldots, n$ and $j = 1, 2, \ldots, m$. ψ and y_j are the activation function and output of the j^{th} node in the hidden layer, respectively.

2) **Output layer**:
After the computation of each output from the neurons in the hidden layer, the information is processed to the output layer. The output layer also has number of neurons which most likely less than the number of neurons in the hidden layer. Neurons in this layer most likely to have linear sigmoid function. The computed output for neurons in the output layer is presented in Equation 3.

$$Y = \varphi(\sum_{j=1}^{k} W_j y_j) \tag{3}$$

k is the number of neurons in the output layers. φ is the linear activation function. Y is the neural network output from the single neuron in the output layer as in our case study.

The learning process continues till we minimize a cost function. In our case, the cost function minimizes the difference between the actual and the result of the network as described in Equation 4. It is defined as the Root Mean Square (RMSE). RMSE can be described by Equation 4.

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{4}$$

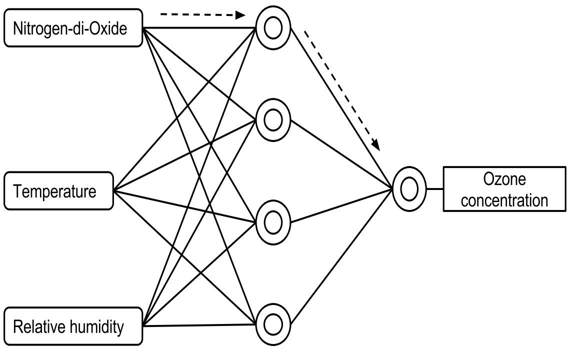


Fig. 1. Feedforward neural network for Ozone prediction

III. Genetic Programming

Genetic Programming is an evolutionary process which was successfully used to solve diversity of problem in system identification and control [19], [20]. GP was inspired from idea of nature selection and evolution introduced by Darwain. GP uses the concept of survival of the fitness to develop solutions that more likely fits to a problem. It is a population based approach. In GP, the population comes in a form of tree structure not a chromosome such as in the case of Genetic Algorithms (GAS) [21]–[24]. A block diagram which shows the GP evolutionary process is presented in Figure 2.

A. Population Initialization and Tree Representation

The initial population for any evolutionary process is produced most likely randomly. In GP a random population P_0 of tress is generated. Each tree represents a solution of a given problem. GP evolves tree structures which is composed of a set of functions and terminals sets provided by the user.

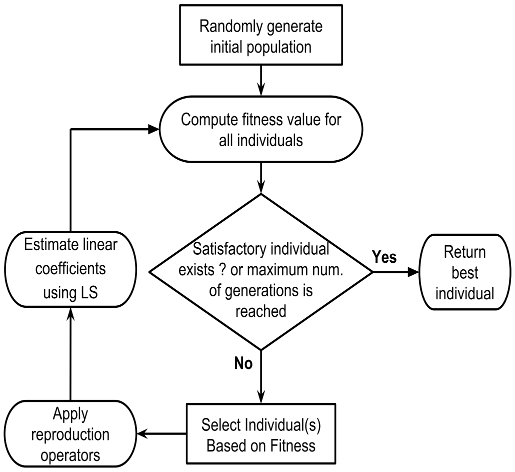

Randomly generate initial population
Compute fitness value for all individuals
Satisfactory individual exists ? or maximum num. of generations is reached
Yes
Return best individual
No
Select Individual(s) Based on Fitness
Apply reproduction operators
Estimate linear coefficients using LS

Fig. 2. Flow chart of the GP technique

The fitness of the initial population is computed according to a given fitness function.

B. Function and Terminal Sets

To develop a mathematical model which represents a relationship between input and output variables, we have to define both function and terminal sets. For a set of inputs $x_1, x_2, x3$ and x_4 to produce an output y, we may have a tree structure which produce the Equation 5. The function ϑ and terminal χ sets are given in Equation 6.

$$\begin{aligned} y &= \zeta(x_1, x_2, x_3, x_4) \\ &= a \times x_1 \times x_2 + b \times \frac{x_3}{x_4} \end{aligned} \tag{5}$$

$$\begin{aligned} \vartheta &= \{\times, +, \div\} \\ \chi &= \{x_1, x_2, x_3, x_4, \theta\} \end{aligned} \tag{6}$$

θ is defined as a random floating point number such that $\theta \in [-1, 1]$. Thus, a and b are also defined in the domain $a, b \in [-1, 1]$

C. Selection Mechanism

While population evolves, selecting individuals for both crossover and mutation depends on what is called the selection mechanism. This is essential process in the generation of new population. Many selection mechanism were presented [25]. They include roulette wheel technique, stochastic universal sampling, tournament selection and many others [26]. Number of selection mechanism used in GP were presented in [27].

D. Multigene Symbolic Regression GP

Symbolic regression method was presented by J. Koza [19]. The objective of this method is to search the space of possible mathematical expressions (i.e. equations) while minimizing some error criteria. Developing mathematical function between input variables x_i and an output y is a challenge. It is important to find the function ζ which relates the inputs and output. Symbolic regression explores both the space of models along with the space of all possible parameters simultaneously such that it can find the best model which minimize the error criterion.

E. Crossover

Crossover is the main operator in any evolutionary process. Crossover is performed between two individuals (i.e. Tree) [28]. A study of crossover operators in GP was presented in [29]. Assuming we have two parents of genes $T_1, \ldots, T_5$ and $R_1, \ldots, R_3$. In Table I, we show the crossover operation in multigene GP.

TABLE I
CROSSOVER IN MULTIGENE GP

T_1	T_2	T_3	T_4	T_5		R_1	R_2	R_3
T_1	R_2	R_3	T_4	T_5		R_1	T_2	T_3

In Multigene symbolic regression, the model output $\hat{y}$ is formed by a weighted output of each of the trees/genes in the multigene individual plus a bias term. Each tree is a function of zero or more of the n inputs variables $x_1, \ldots, x_n$. Mathematically, a Multigene regression model can be written as:

$$\hat{y} = \delta_0 + \delta_1 \times Tree_1 + \cdots + \delta_m \times Tree_m \tag{7}$$

δ_0 represents the bias or offset term while $\delta_1, \ldots, \delta_m$ are the gene weights and m is the number of genes (i.e. trees) which constitute the available individual. The values of δ coefficients can be estimated using least square estimation technique. A simple example of two multigene model is presented in Figure 3 and Equation 8.

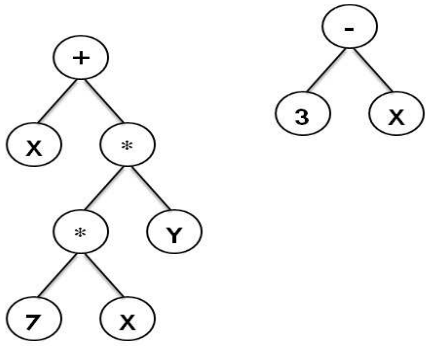


Fig. 3. Example of a Multigene Symbolic GP model

$$\delta_0 + \delta_1[X + (7 \times X) \times Y] + \delta_2[3 - X] \tag{8}$$

F. Mutation

Mutation is a relatively important operator it helps in keeping diversity in the population especially when most individual has the same fitness. Mutation helps keeping the exploration in the population. Mutation in multigene GP operates almost the same way as in standard GP.

IV. Performance Criterion

Number of performance criterion were used to evaluate the performance of the developed ANN and Multigene GP models. These evaluation criterion are presented in the following equations.

1) Euclidian distance (ED):

$$ED = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{9}$$

2) Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{10}$$

3) Mean Magnitude of Relative Error (MMRE):

$$MMRE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \tag{11}$$

where y and $\hat{y}$ are the actual measured Ozone level and the predicted Ozone level developed by the ANN and GP models given n measurements.

V. Site characterization and data

The study area under study is Chenbagaramanputhur. It is a rural place in Kanyakumari district and is about 12 km from Nagercoil town. In the North and North East of the city, you can find the Tirunelveli district. Kerala State is located in the North West and sea in the west and south of Chenbagaramanputhur (See Figure 4).

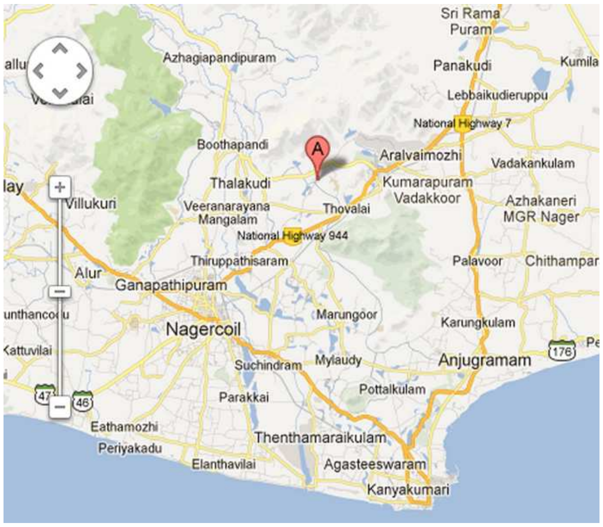


Fig. 4. Location of the area of study at Chenbagaramanputhur

The data used in this study were reported in [13]. Authors in [13] mentioned that the measurements were collected using a portable Aeroqual series S200. The Aeroqual series 200 can measure various ozone levels. Measurements were taken every 3 hours intervals for a period of 3 months during May 2009 to July 2009. Figure 5 shows the inputs and output of the proposed models. The variables used as inputs and output are presented in Table II.

TABLE II
Inputs and Output model variables

Inputs	Nitrogen dioxide concentration	x_1
	Mean temperature	x_2
	Prevailing % Relative Humidity	x_3
Output	Mean surface ozone concentration O_3	y

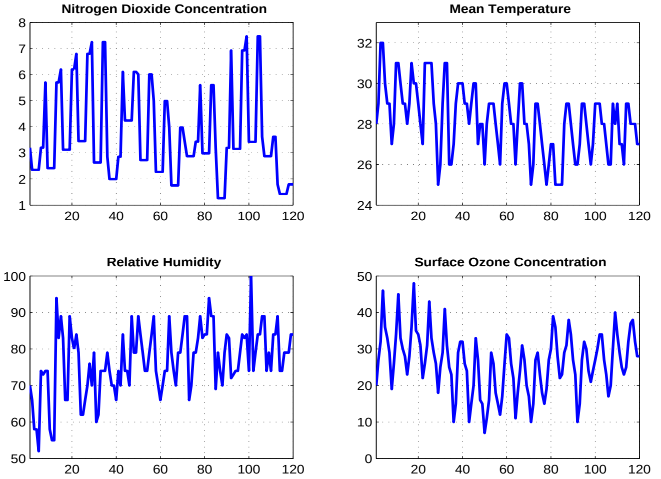


Fig. 5. Training, Testing and Validation data set [13]

VI. Experimental Setup and Results

A. Developed MLP-ANN Model

We developed a MLP-ANN model using the input-output data presented in Table II to model the surface Ozone using the parameters given in Table III. Various number of neurons in the hidden layer were explored during the learning process. The best number of neurons found was four. Figure 6, shows that the MLP training process had fast convergence to the minimum training error after only nine cycles (epchs). Figure 6 shows the actual and estimated Ozone surface values based the final developed MLP model.

B. Developed Multigene GP Model

To develop the genetic programming model, GPTIPS Matlab Toolbox developed in [28] is used. GPTIPS is a powerful genetic programming software tool which can be used of modeling of dynamical nonlinear systems. The tool can be configured to evolve multigene tree structure. The Multigene approach often develops simpler models than evolving models consisting of one monolithic GP tree.

The data set described earlier was loaded then the Multigene GP was applied using GPTIPS Tool. The parameters of

TABLE III
Neural Network parameters

Parameter	Value
Architecture	Multi Layer perceptron
Number of hidden layers	1
Nodes in first hidden layer	4
Epochs	50
Training method	Scaled conjugate gradient

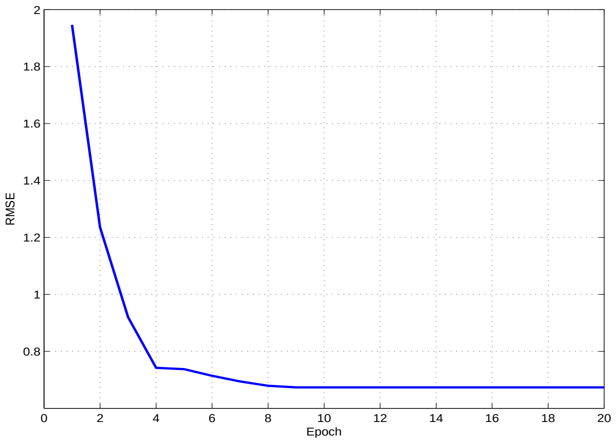


Fig. 6. Convergence of the MLP-ANN

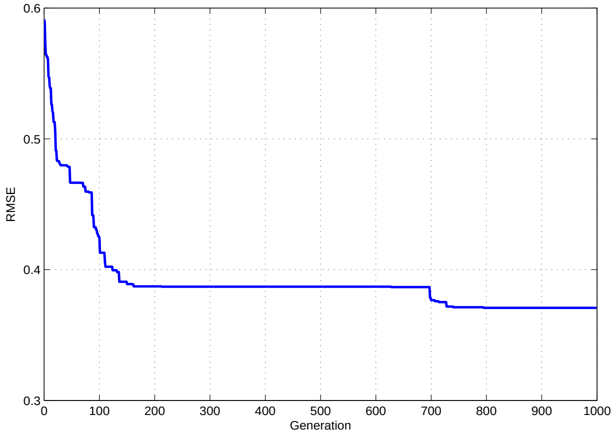


Fig. 8. Convergence of the GP evolutionary process

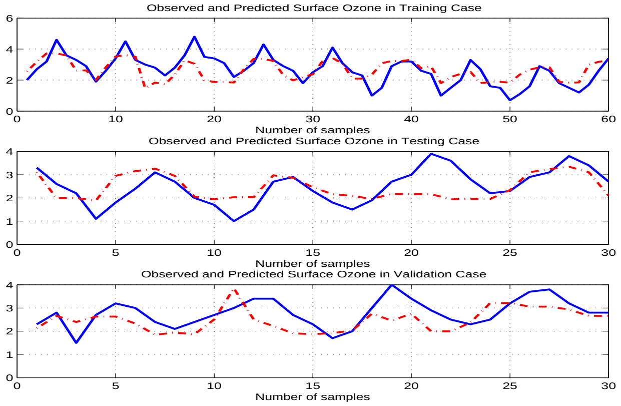


Fig. 7. Observed and Predicted O_3 using MLP-ANN model

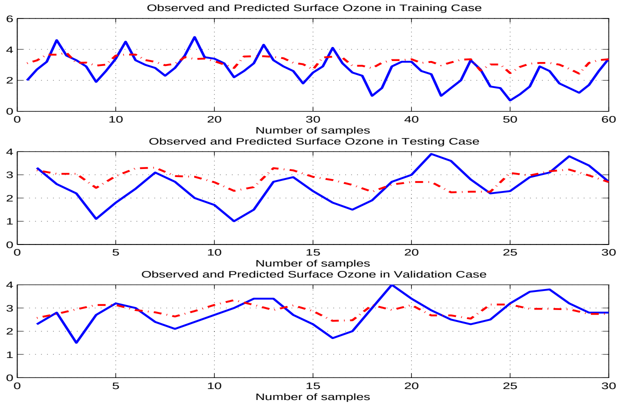


Fig. 9. Observed and Predicted O_3 using Multigene GP Model

the algorithm were tuned as listed in Table IV. In Figure 8, the convergence of GP over 100 generations is shown. The best generated Surface Ozone Multigene GP model is given in Equation 12. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. Figure 9 shows the actual and estimated surface Ozone values based the developed GP model.

TABLE IV
GP TUNING PARAMETERS

Population size	50
Number of generation	1000
Selection mechanism	Tournament
Max. tree depth	12
Probability of Crossover	0.85
Probability of Mutation	0.1
Max. No. of genes allowed in an individual	7

$$\begin{aligned} y &= 0.01442\, x_1^2 - 0.9507\, x_2 - 0.2634\, x_3 - 0.7902\, x_1 \\ &+ 0.006796\, x_2^2 + 0.000828\ \ x_3^2 + 0.03329\, x_1 x_2 \\ &- 0.0001513\, x_2^2\, (2x_1 - x_3) + 30.63 \qquad (12) \end{aligned}$$

C. Comments on Results

In order to compare the performance of GP and MLP for predicting Ozone concentrations, the evaluations criteria discussed in Section IV are used to assess both developed model. The criteria measurements for the models are computed and summarized in Table V. It can be noticed that the Multigene GP model has shown better prediction results over the MLP model for training, testing and validation partitions by means of all evaluation criteria. Moreover, the final developed GP model shown in Equation 12 is considered much simpler than the complex model of the ANN approach.

VII. CONCLUSIONS AND FUTURE WORK

A comparison between genetic programming and multilayer perceptron neural networks were presented for short term prediction of surface Ozone based on limited number of measured pollutant and meteorological variables. The GP approach adopted is based on Multigene symbolic regression which generates mathematical models of linear combinations of low order non-linear transformations of the input variables. Based on this comparison, it can be concluded that the evolutionary models of the Multigene GP have promising potential for predicting surface ozone concentrations when

TABLE V
EVALUATION CRITERIA FOR THE DEVELOPED MODELS

	Multigene GP			MLP-ANN		
	Training	Testing	Validation	Training	Testing	Validation
RMSE	0.90342	0.75708	0.51826	0.74203	0.68969	0.58236
ED	6.9979	4.1467	2.8386	5.7477	3.7776	3.1897
MAE	0.71996	0.62665	0.40887	0.59362	0.53972	0.48025
MMRE	0.414	0.32339	0.16539	0.29577	0.24084	0.19439

the available number of measured pollutant and meteorological variables is limited as the case investigated in this study. The Ozone Multigene GP model was also a compact model. Future investigation on Multigene GP and other soft computing techniques on handling the environmental monitoring problems will be considered.

ACKNOWLEDGEMENT

This research is funded by the Deanship of Research at Zarqa University, Jordan.

REFERENCES

[1] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, and E. Alper, "Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach," *Advances in Environmental Research*, vol. 5, no. 1, pp. 47 – 59, 2001.

[2] S. Abdul-Wahab and S. Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," *Environmental Modelling & Software*, vol. 17, no. 3, pp. 219 – 228, 2002.

[3] M. Norval, R. M. Lucas, A. P. Cullen, F. R. de Gruijl, J. Longstreth, Y. Takizawa, and J. C. van der Leun, "The human health effects of ozone depletion and interactions with climate change," *Photochem. Photobiol. Sci.*, vol. 10, pp. 199–225, 2011.

[4] H. Trker and M. Yel, "Effects of ultraviolet radiation on mole rats kidney: A histopathologic and ultrastructural study," *Journal of Radiation Research and Applied Sciences*, 2014.

[5] M. Bell, R. Goldberg, C. Hogrefe, P. Kinney, K. Knowlton, B. Lynn, J. Rosenthal, C. Rosenzweig, and J. Patz, "Climate change, ambient ozone, and health in 50 us cities," *Climatic Change*, vol. 82, pp. 61–76, 2007.

[6] B. Carbon, "Monitoring of CO, NO2, SO2, ozone, benzene and benzo(a)pyrene in new zealand, air quality technical report no. 42," Tech. Rep., 2004.

[7] O. Pastor-Bárcenas, E. Soria-Olivas, J. D. Mart'ın-Guerrero, G. Camps-Valls, J. L. Carrasco-Rodr'ıguez, and S. del Valle-Tascón, "Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modelling," *Ecological Modelling*, vol. 182, no. 2, pp. 149–158, 2005.

[8] E. Agirre, A. Anta, and L. J. R. Barron, "Forecasting ozone levels using artificial neural networks," *Forecasting Models*, pp. 208–218, 2010.

[9] V. R. Prybutok, J. Yi, and D. Mitchell, "Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentrations," *European Journal of Operational Research*, vol. 122, no. 1, pp. 31 – 40, 2000.

[10] V. Gvozdic, E. Kovac-Andric, and J. Brana, "Influence of meteorological factors NO2, SO2, CO and PM10 on the concentration of O3 in the urban atmosphere of eastern croatia," *Environmental Modeling and Assessment*, vol. 16, 2011.

[11] N. Banan, M. T. Latif, L. Juneng, and M. F. Khan, "An application of artificial neural networks for the prediction of surface ozone concentrations in malaysia," in *From Sources to Solution*. Springer, 2014, pp. 7–12.

[12] H. Faris, M. Alkasassbeh, and A. Rodan, "Artificial neural networks for surface ozone prediction: Models and analysis." *Polish Journal of Environmental Studies*, vol. 23, no. 2, 2014.

[13] R. S. Selvaraj, K. Elampari, R. GAYATHRI, and S. J. JEYAKUMAR, "A neural network model for short term prediction of surface ozone at tropical city," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5306–5312, 2010.

[14] A. Sheta, N. Ghatasheh, and H. Faris, "Forecasting global carbon dioxide emission using auto-regressive with exogenous input and evolutionary product unit neural network models," in *Information and Communication Systems (ICICS), 2015 6th International Conference on*, April 2015, pp. 182–187.

[15] M. Alkasassbeh, A. F. Sheta, H. Faris, and H. Turabieh, "Prediction of pm10 and tsp air pollution parameters using artificial neural network autoregressive, external input models: A case study in salt, jordan," *Middle-East Journal of Scientific Research*, vol. 14, no. 7, pp. 999–1009, 2013.

[16] H. Faris and A. Sheta, "Identification of the tennessee eastman chemical process reactor using genetic programming," *International Journal of Advanced Science and Technology*, vol. 50, pp. 121–140, Jan. 2013.

[17] A. Sheta and R. Hiary, "Modeling lipase production process using artificial neural networks," in *Proceedings of the 3rd IEEE International Conference on Multimedia Computing and Systems*, Tangier, Morocco, 10-12 May 2012, pp. 1158–1163.

[18] A. Sheta and M. El-Sherif, "Optimal prediction of the nile river flow using neural networks," in *Proceedings of the International Joint Conference on Neural Networks, Washington, D.C., July*, vol. 5, 1999, pp. 3438–3441.

[19] J. Koza, "Evolving a computer program to generate random numbers using the genetic programming paradigm," in *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann, La Jolla,CA, 1991.

[20] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[21] K. A. De Jong, "Adaptive system design: A genetic approach," *IEEE Transaction Sys. Man. Cybern.*, vol. 10, no. 3, pp. 556–574, 1980.

[22] K. De Jong, "Are genetic algorithms function optimizers?" in *Proceedings of the Second Parallel Problem Solving From Nature Conference*. The Netherlands: Elsevier Science Press, 1992, pp. 3–14.

[23] A. Sheta and K. D. Jong, "Parameter estimation of nonlinear systems in noisy environment using genetic algorithms," in *Proceedings of the IEEE International Symposium on Intelligent Control (ISIC'96)*, 1996, pp. 360–366.

[24] A. Sheta and K. De Jong, "Time-series forecasting using GA-tuned radial basis functions," in *Information Science Journal*, 2001, pp. 221–228.

[25] B. L. Miller, B. L. Miller, D. E. Goldberg, and D. E. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex Systems*, vol. 9, pp. 193–212, 1995.

[26] S. Legg, M. Hutter, and A. Kumar, "Tournament versus fitness uniform selection," in *Proc. 2004 Congress on Evolutionary Computation (CEC-2004)*. Portland, OR: IEEE, 2004, pp. 2144–2151.

[27] E. Galvan-Lopez, B. Cody-Kenny, L. Trujillo, and A. Kattan, "Using semantics in the selection mechanism in genetic programming: A simple method for promoting semantic diversity," in *Proceedings of the 2013 IEEE Congress on Evolutionary Computation (CEC)*, June 2013, pp. 2972–2979.

[28] D. P. Searson, D. E. Leahy, and M. J. Willis, "GPTIPS : An open source genetic programming toolbox for multigene symbolic regression," in *Proceedings of the International Multi-conference of Engineers and Computer Scientists 2010 (IMECS 2010)*, vol. 1, Hong Kong, 17-19 Mar. 2010, pp. 77–80.

[29] W. Spears and V. Anand, "A study of crossover operators in genetic programming," in *Methodologies for Intelligent Systems*, ser. Lecture Notes in Computer Science, Z. Ras and M. Zemankova, Eds. Springer Berlin Heidelberg, 1991, vol. 542, pp. 409–418.

Video conference Android platform by your mobile phone

Mr. Mohamed Khalifa
WIMCS Research Team, ENET'COM
Sfax-University, Tunisia

Dr. Chaafa Hamrouni
Telecommunication department
CERT
Sfax, Tunisia

Pr. Mahmoud Abdellaoui
WIMCS Research Team, ENET'COM
Sfax-University, Tunisia

***Abstract*— Video conferencing is a visual and audio communications technology dedicated to Smartphones. It is based on the client-server communication model that this requires many limits since using the server. In this paper, we proposed a new communication model without going through the server (client-client), in addition, this model makes video conference used by mobile multimedia phones.**

Keywords—videoconference; multimedia mobile phone; streaming; Smartphone; Server; Customers.

I. Introduction

On 1983, the first classic mobile phone is launched by Motorola; it transmits and receives text messages and calls. It often has a rudimentary camera and may be a game or two. After he was replaced by the media mobile phone that has many additional functions, made possible through the integration of an advanced operating system in the phone. The use of these two types of mobile phone gradually decreased to 1992. Since 1994, new cells phone categories are offered to the public: Smartphone that supported many functions of a PDA and a laptop. The use of these handheld devices are exploded in the world of mobile telephony, resulting in partial loss of other types of mobile phones even if they are expensive, subject to theft, are not within the reach of all world, easily breakable... In addition, a comparison was made between the mid-phones and multimedia mobile phones with the operating system level, memory, video and audio format, display protocol... The results of this comparison show that differences and characteristics between these two categories of cell phones are a little more subtle, and then they allow us to make the competitive multimedia cell phones to other ordis-phones. Similarly, these simple lightweight cell phones are intermediate between the Smartphone mobile and conventional phones. They have simplified and various multimedia features, a very good battery life, ease of use, much less fragile than the Smartphone and especially that mobile search is within the reach of everyone. But, they are less popular compared to Smartphones because of these applications. Faced with this gap between the two in uses and applications, you have to technically modify these multimedia mobile phones with the goal to adapt them and bring them level with Smartphone. For thus, in this article we created a usable application in the field of Streaming (video conferencing, broadcasting, remote monitoring, remote training ...) and make it adaptable to multimedia mobile phones. This application allowed transmitting video and audio data between two or more customers without using server. The created application, based on a new model audio and video communication without server, presented a good solution to resolve the weaknesses environment client-server such high cost, limited support, a weak link...

II. Theoretical Analysis

In this part, we have presented, at first, a video-conference constraints and secondly, the operation of this application with and without server.

A. Videoconference constraints

At a meeting in the company headquarters, the presence of all heads of agencies, which are scattered throughout the country, present difficulties; for that, they are obliged to perform a remote meeting through video conferencing. The latter is a widely used technology that allows two (or more) persons to enter remote visual and audio communication and work together (or organizing business meetings or conferences or distance learning courses). So, videoconferencing is the improvement of communications source of increased productivity, reduced costs, time lost from deletion, environmental conservation…. This technique requires a rate sufficient to provide transportation of the picture and sound and also three basic materials such as: microcomputer to display the picture, webcam or camera to capture the picture and microphone for sound. Most video-conferencing applications using a server, is high cost. For that, main goal of this article is the establishment of a videoconference Android platform across your multimedia mobile phone to achieve a "**client-client**" communication model without server.

B. Application with using server

There are two types of Streaming server: one standard server and the other specialized server. In this application, we used the dedicated server since it supports RTSP protocol, against the standard server supports HTTP. Similarly, most multimedia mobile phones support RTSP / RTP. For thus, we chose the RTSP protocol to control the delivery of video and audio data in real time with properties and RTP to transmit multimedia data in real time [1], [2]. The used server in this application is: "Wowza Media Server" [3].

RTP works always with its companion RTCP to allow monitoring of data delivery in a manner scalable to large multicast networks. UDP, that's the best underlying protocol, used for these two protocols as most multimedia mobile phones support RTSP / RTP non-interlaced (RTP over UDP) [4]. To transmit video and audio data in the network, it's best to encode the data with a low bit rate and low complexity encoding. For that, we use the H.263 encoder for the video data and the encoder AMRNB for audio data as most multimedia mobile phones support these encoders. After the selection of appropriate protocols and configuring encoding settings that are adaptable to the capabilities of multimedia mobile phones, we started to have how to build video and audio data and send them into the network. The construction of multimedia data is done in two classes, one for video and one for audio data using the same object "**MediaRecorder**" which is belongs to the super class "**Video_Audio_Data**".

```
mMediaRecorder = new MediaRecorder();
mMediaRecorder.setCamera(camera);
mMediaRecorder.setVideoSource(MediaRecorder.VideoSource.
CAMERA);
mMediaRecorder.setOutputFormat(MediaRecorder.OutputFormat.
THREE_GPP);
mMediaRecorder.setVideoEncoder(MediaRecorder.VideoEncoder.
H263);
mMediaRecorder.setPreviewDisplay(mSurfaceView.getHolder().
getSurface());
mMediaRecorder.setVideoSize(176,144);
mMediaRecorder.setVideoFrameRate(15);
mMediaRecorder.setVideoEncodingBitRate((int)(170*0.8));
mMediaRecorder.setOutputFile((LocalSocket)emetteur
.getFileDescriptor());
mMediaRecorder.prepare();     mMediaRecorder.start();
```

Fig. 1. Video data conception program.

```
mMediaRecorder = new MediaRecorder();
mMediaRecorder.setAudioSourc(MediaRecorder.AudioSource.CAMCORD
ER);
mMediaRecorder.setOutputFormat(MediaRecorder.OutputFormat.AMR_NB
);
mMediaRecorder.setAudioEncoder(MediaRecorder.AudioEncoder.AMR_N
B);
mMediaRecorder.setAudioChannels(2);
mMediaRecorder.setAudioSamplingRate(8000);
mMediaRecorder.setAudioEncodingBitRate(16000);
mMediaRecorder.setOutputFile((LocalSocket)emetteur.getFileDescript());
mMediaRecorder.prepare();  mMediaRecorder.start();
```

Fig. 2. Audio data conception program.

After the multimedia data establishment, we transferred these to the "**RTP_RTCP_Paquet**" class that can fill data camera and microphone in a RTP packet from the length of the header of the packet to the value MTU in the format of the header of RTP packet payload to use for AMRNB (audio) and H263 (video) using the class "**Data_AMRNB**" and class "**Data_H263**" in order to create an RTP packet that will be sent in unicast or multicast and an RTCP packet that will be sent every three seconds [5], [6].

```
RTP_RTCP_Paquet.setInputStream((LocalSocket)recepteur.
getInputStream());
RTP_RTCP_Paquet.start();
```

Fig. 3. Creat packet

Before starting the dissemination of multimedia data in real time, RTSP streams are exchanged between client-server and server-client to provide customers a URL can read the media on the server. These exchange messages are described by the following figure:

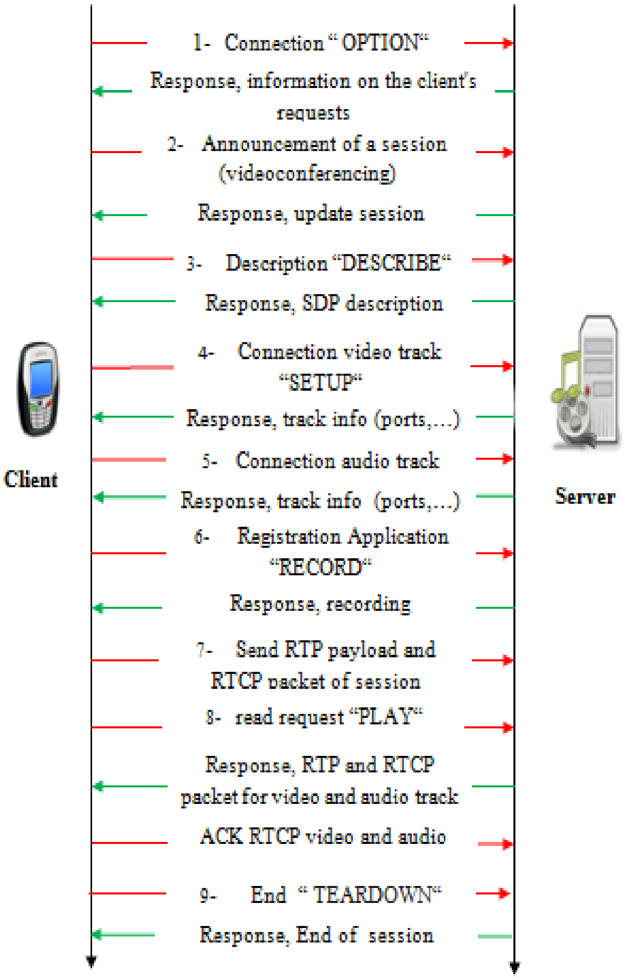

1- Connection " OPTION"
Response, information on the client's requests
2- Announcement of a session (videoconferencing)
Response, update session
3- Description "DESCRIBE"
Response, SDP description
4- Connection video track "SETUP"
Response, track info (ports,...)
5- Connection audio track
Response, track info (ports,...)
6- Registration Application "RECORD"
Response, recording
7- Send RTP payload and RTCP packet of session
8- read request "PLAY"
Response, RTP and RTCP packet for video and audio track
ACK RTCP video and audio
9- End " TEARDOWN"
Response, End of session
Client
Server

Fig. 4. Diagram of exchanges between the server and the client.

1. The client initiates the session with the server by sending a request "OPTIONS". The server responds to this request with information on what to support and what kind of requests it can receive.
2. After launching session, the client announces the session type: video conferencing, television ... The server responds with the session update.

3. The client sends a message describing. The server responds with an SDP file that the client can use to get more information about the content that will be sent by the server [7].

4. The client sends a setup request to the server to the video track at the end to inform the server that it will use UDP ports for RTP and RTCP communication. The server responds with information about the UDP ports that will be used by the server for this session.

5. The client sends a setup request to the server to the video track, the end to inform the server that it will use UDP ports for RTP and RTCP communication. The server responds with information about the UDP ports that will be used by the server for this session.

6. After the establishment of the communication ports, the client requests to record the session. The server saves the session to become accessible by other customers.

7. After the recording session, the client begins sending the RTP stream of the session.

8. The client sends a read request that informs the server that it is ready to receive the RTP data stream. The server starts sending the RTP payload and RTCP packet, also, the client sends RTCP packet to the server.

Here is a sample source code which describes how to read media data located on the server with the subject "MediaPlayer":

```
player.setAudioStreamType(AudioManager.STREAM_MUSIC);
player.setDataSource("rtsp://@IP_de_serveur:port_serveur/live/
test.stream");
player.setOnErrorListener(this);
player.setOnPreparedListener(this);
player.prepareAsync();
player.start();
```

Fig. 5. Read media data.

9. If the client closes the read a "TEARDOWN" request is sent to the server so it stops the streaming session.

C. *Application without using server*

In this part, we used the same encoders for video and audio data, but we only used the RTP and UDP transport protocols. In addition, we have optimized the messages exchanged between the client and the server to allow a multimedia cell phone to play the same role server to perform a visual and audio communication between two (or more) multimedia mobile phones without going through the server. This communication is based on the SDP. Hence session description record in the internal memory of the mobile device as a text file and read this file directly using a media player (VLC, FFmpeg, ...). These optimized exchanges messages are described by the following figure:

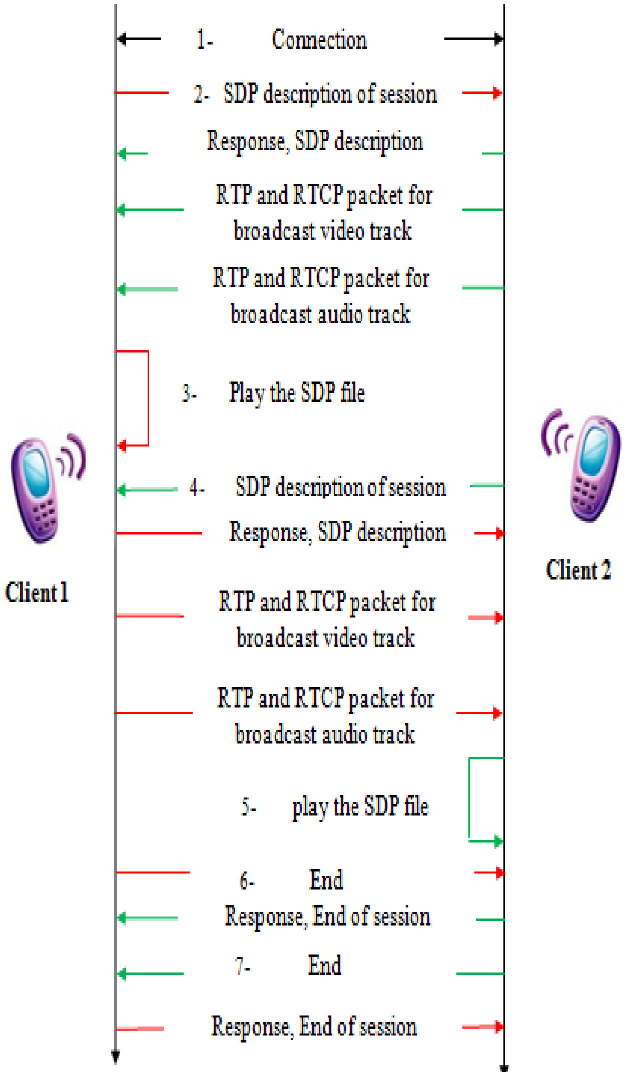

1- Connection
2- SDP description of session
Response, SDP description
RTP and RTCP packet for broadcast video track
RTP and RTCP packet for broadcast audio track
3- Play the SDP file
4- SDP description of session
Response, SDP description
RTP and RTCP packet for broadcast video track
RTP and RTCP packet for broadcast audio track
5- play the SDP file
6- End
Response, End of session
7- End
Response, End of session
Client 1
Client 2

Fig. 6. Diagram of exchanges between clients without server.

Before starting the session, the connection establishment between the two clients is done reliably using TCP / IP to transmit the SDP description. For cons, the RTP payload transmission and the report of the sender is unreliably using RTP over UDP. This is explained by the following steps:

1. Establishing connection.

2. After the connection establishment, the client 1 demands SDP description of the client 2. The latter responds with an SDP text file. The client 1 will also receive the RTP payload and RTCP packet.

3. The client 1 stores the SDP description in an internal memory as a text-file and then read that file.

```
v=0
o=- 0 0 IN IP4 null
s=Unnamed
i=N/A
c=IN IP4 192.168.168.103
t=0 0
a=recvonly
m=audio 5004 RTP/AVP 96
a=rtpmap:96 mpeg4-generic/16000
a=fmtp:96 streamtype=5; profile-level-id=15;mode=AAC-hbr;
config=1408; SizeLength=13; IndexLength=3; IndexDeltaLength=3;
a=control:trackID=0
m=video 5006 RTP/AVP 96
a=rtpmap:96 H263-1998/90000
a=control:trackID=1
```

Fig. 7. Description of SDP

4. Similarly, the client 2 requests the SDP description. The client 1 responds with a text file also it will send the RTP payload and RTCP packet.

5. The client 2 stores the SDP description in an internal memory as a text-file and then read that file.

6. If the client 1 closes the playback. The client 2 also closes the streaming session.

7. If the client 2 closes the playback. The client 1 also closes the streaming session.

This application is not limited only between devices but it can be used by a group of multimedia mobile phones. Hence, every customer of the group sends the session description to a group of receivers. This group is specified by a multicast IP address (224.0.0.0 to 239.255.255.255) [8].

III. Study of Cases

In this part, we studied two cases of this application, one with server and another without server.

A. *Case study with using server*

This application allows a user to open two sessions (video session and audio session):

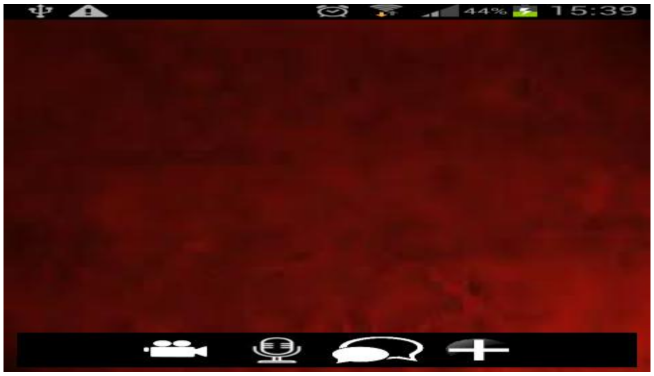


Fig. 8. Home application.

If a user wants to open a video session he has to press the image "camera". Hence, the video session activity will be created:

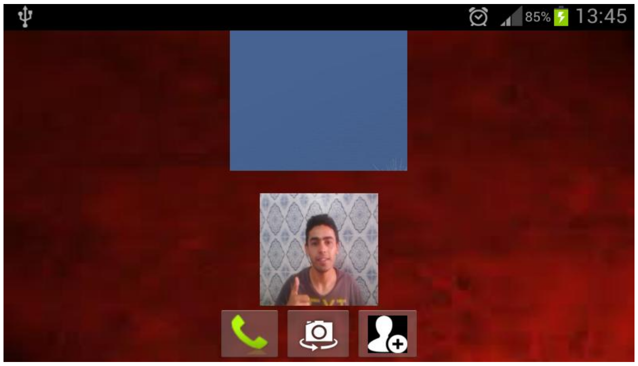


Fig. 9. Activity video session.

If the user wants to establish a connection with the server, he must press the "call" image. So, a dialog box will be displayed to enter the authentication settings for the server and select the video quality settings:

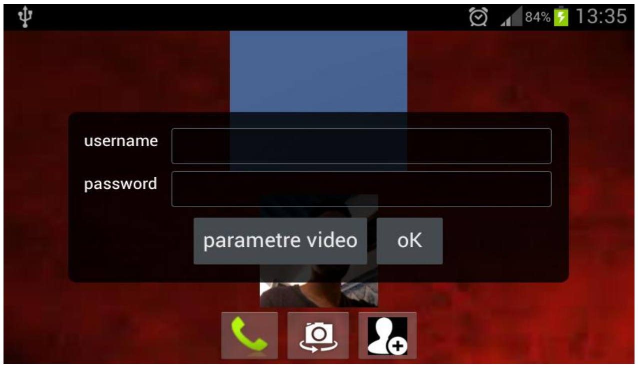


Fig. 10. Box authentification.

To connect to the server, the user must press the "OK" button. After the connection establishment, the server saves the session of the client and it will also receive data from camera and microphone for this session. If the user wants to open a visual and audio communication, it must press the "pseudo" image to enter the nickname of the other participant:

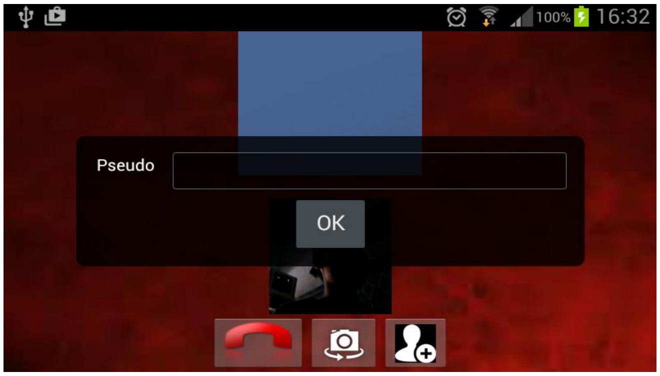


Fig. 11. Pseudo.

After adding nickname, the user must press the "OK" button to perform a visual and audio communication:

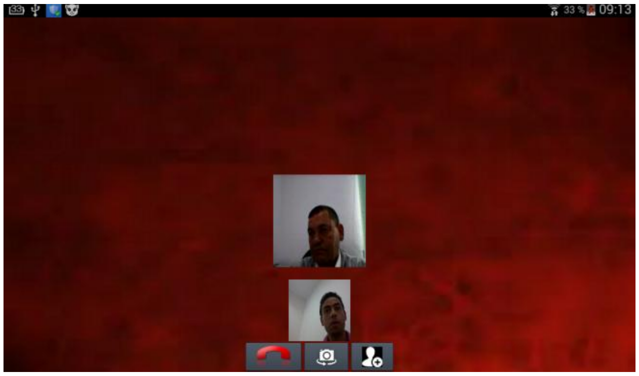


Fig. 12. Transmitter side.

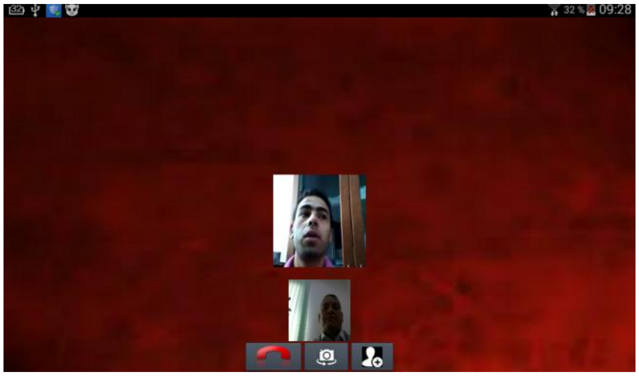


Fig. 13. Receiver side.

B. *Case study without using server*

After creating the activity video session, the user must press on the image "call" to enter the identifiers of the other Participant:

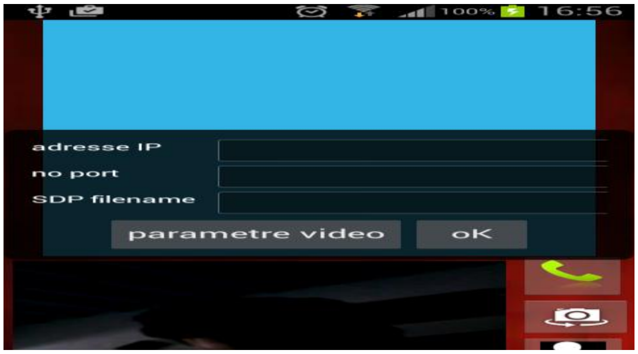


Fig. 14. Connection parameter.

After the connection establishment, a text file containing the SDP description will be stored in the internal memory of each client. For now, in this application reading the file is through a media player like VLC to make a visual and audible communication:

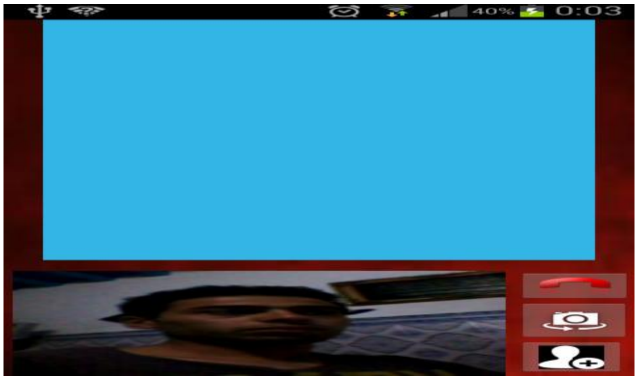


Fig. 15. Transmitter side.

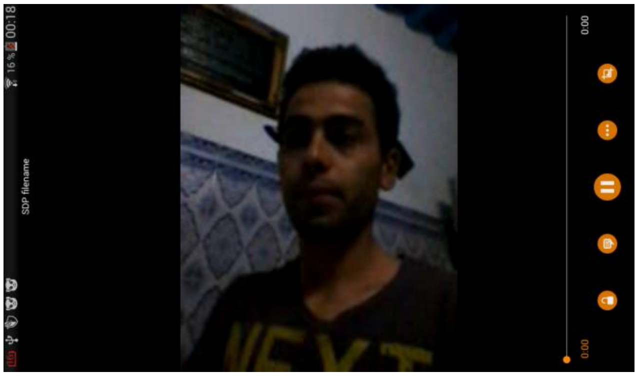


Fig. 16. Read SDP file with VLC.

IV. Conclusion

In this work, we have established a new model of video and audio communication from multimedia mobile phone without going through the server. This model is more efficient than the existing model at the level of simplicity, cost, use... The advantages offered by the implementation of this model are manifested in adaptation capabilities of all mobiles that contain three basic material video conferences (computer, camera and microphone).

Abbreviations and Acronyms

PDA: Personal Digital Assistant; RTSP: Real Time Streaming Protocol; HTTP: Hyper Text Transfer Protocol; RTP: Real-time Transport Protocol; RTCP: Real-time Transport Control Protocol; UDP: User Datagram Protocol; H.263: Video Codec; AMRNB: Adaptive Multi-Rate Narrow-Band; MTU: Maximum Transfer Unit; URL: Uniform Resource Locator; SDP: Session Description Protocol; TCP: Transmission Control Protocol; IP: Internet Protocol.

ACKNOWLEDGMENT

This work has been accomplished at WIMCS-Team research, ENET'COM-Sfax University. Part of this work has been supported by CERT-Tunisia Telecom and MESRSTICS Scientific Research Group-Tunisia.

REFERENCES

[1] R. Lanphier and H. Schulzrinne, "Real Time Streaming Protocol", IETF, REF 2326, Avril 1998, pp. 92.

[2] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson "RTP: A Transport Protocol for Real-Time Applications", Audio-Video Transport Working Group, REF 1889, January 1996, pp. 75.

[3] The Wowza Media Server. Available http://www.wowza.com/pricing/installer.

[4] J. Postel, "User Datagram Protocol", IETF, REF 768, 28 august 1980, pp. 3.

[5] J. Sjoberg, M. Westerlund, A. Lakaniemi and Q. Xie, "Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-Rate (AMRNB) and Adaptive Multi-Rate Wideband (AMRWB) Audio Codec, IETF, REF 3267, june 2002, pp. 49.

[6] J. Ott, C. Bormann, G. Sullivan, S. Wenger and R. Even, "RTP Payload Format for ITU-T Rec. H.263 Video", IETF, REF 4629, January 2007, pp. 29.

[7] M. Handley, V. Jacobson, "SDP: Session Description Protocol", IETF, REF 2327, April 1998, pp. 42.

[8] S. Deering and Stanford University, "Host Extensions For IP Multicasting", IETF, REF 1054, May 1988, pp. 19.

Implementation of Vision-based Object Tracking Algorithms for Motor Skill Assessments

Beatrice Floyd and Kiju Lee
Department of Mechanical and Aerospace Engineering
Case Western Reserve University
Cleveland, Ohio 44106

***Abstract*—Assessment of upper extremity motor skills often involves object manipulation, drawing or writing using a pencil, or performing specific gestures. Traditional assessment of such skills usually requires a trained person to record the time and accuracy resulting in a process that can be labor intensive and costly. Automating the entire assessment process will potentially lower the cost, produce electronically recorded data, broaden the implementations, and provide additional assessment information. This paper presents a low-cost, versatile, and easy-to-use algorithm to automatically detect and track single or multiple well-defined geometric shapes or markers. It therefore can be applied to a wide range of assessment protocols that involve object manipulation or hand and arm gestures. The algorithm localizes the objects using color thresholding and morphological operations and then estimates their 3-dimensional pose. The utility of the algorithm is demonstrated by implementing it for automating the following five protocols: the sport of Cup Stacking, the Soda Pop Coordination test, the Wechsler Block Design test, the visual-motor integration test, and gesture recognition.**

Keywords*—*Vision-based Object Tracking; Motor Skill Assessment; Multi-marker Tracking; Computer-based Assessment.

I. Introduction

Assessment of upper extremity motor skills often involves manipulating physical objects, hand drawing and writing, or performing specific gestures [1] - [7]. Early assessment of such skills can potentially lead to early diagnosis of any deficits and thus result in better treatment outcomes in the long term [1], [8]. For example, motor skills deficiencies can be observed and are symptomatic of a learning or developmental disability, a traumatic brain injury, and normal aging. Assessment of such skills by a human clinician may encounter several challenges such as high cost [9]; time constraints [3]; and inconsistent professional awareness and expertise in diagnosis [10], [11]. Advancement in computing and sensing technologies have enabled automation of such assessment tasks previously conducted by human administrators. Automation does not only improve the accuracy and efficiency of tasks, but also can accomplish tasks that were previously impossible using human skills alone [12]. The assessment of motor skills would in particular benefit from automation. This is partly because of the increased accuracy, efficiency, and consistency of the measurements, but more notably automation can result in quantitative information that would not be possible from traditional manual assessment methods. For example, the Box and Block Test of Manual Dexterity (BBT) could be automated by installing RF readers in the two boxes and embedding RF tags in all the blocks [13], [14]. The system was able to automatically sense when the blocks were placed in either of the boxes based on the relative signals from the two readers [15], [16]. This resulted in the same assessment data as the manual assessment while being more time efficient and collecting more data about the blocks movements.

Automation of upper extremity motor skill assessment that involves object manipulation can be realized in two ways: i) by employing *active* objects with embedded sensing and communication capabilities or ii) using *passive* objects with an external sensing device(s). It may be a combination of the two. Over the past decade, a variety of active objects have been developed for a broad range of education, entertainment, and research purposes [17]. Several studies have used the sensor-embedded blocks for measuring three-dimensional (3D) spatial cognitive abilities by observing construction patterns and performance [18]-[20]. Learning Block is a digitally augmented physical block system enriched with a speaker and LED display [21]. It aims to function as a playful learning interface for children via embedded gesture recognition. Another interesting application is the use of a sensor-embedded block system, called Navigation Blocks, for tangible navigation of digital information through tactile manipulation and haptic feedback [22]. Tangibles is also an active object system designed for tangible manipulation and exploration of digital information [23]. There are also block systems integrated with sound feedback. For example, AudioBlocks and Block Jam features an augmented sound feedback mechanism to enable users to design musical sequences by manipulating the tangible objects with visual and sound feedback [24], [25]. Multi-agent autonomous interactive blocks and games were developed specifically for behavioral training of children with an autism spectrum disorder [26].

Most of the existing work on object manipulation and gesture detection using passive objects has been geared toward vision-based approaches [27], [28]. For example, a depth-sensing camera was used to build a height map of the objects on an interactive tabletop platform for recognizing objects and detecting interaction between the player and the objects [29]. PlayAnywhere is a projection-vision system that can detect hover and touch by a human finger on a tabletop with a projected image [30]. Another interesting system is called Touch-Space, which is a game environment that combines reality with a virtual game environment based on ubiquitous, tangible, and social computing [31], [32]. Vision-based systems, compared to the methods using active objects, allow flexibility in the game or test design and the types of applications. However, most of these algorithms are computationally expensive [30]. In addition, sensing is limited to the vision range unless additional sensing devices are used. Using active objects with embedded sensors or combining the two approaches may overcome the limitations of a vision-only method, but the hardware can be costly, in particular if a large number of objects are employed, and it is difficult to make a versatile method due to inflexibility of the hardware [33], [34].

This paper presents an algorithm designed for assessing object manipulation skills and hand gestures using a single standard webcam. No additional equipment other than a webcam is required. The algorithm is based on color thresholding for initial localization and morphological operations to find the object's edges. The corners are then identified by transferring the edges and used for pose estimation in real time. The result is the three-dimensional (3D) pose of the object which can be used for test automation and additional behavioral assessments. The algorithm described here is for tracking well-defined objects or markers rather than directly tracking hands and arms to simplify the computational complexity. Tracking hand motions would give a lot of interesting information about the person's upper extremity motor skills as explored by other researchers [35]. However, it is not necessary or ideal for object-based motor skill assessments for several reasons. First, the assessments being automated do not rely on hand position information, but instead on the resulting position of objects. Thus it would be counterproductive to track the hands since it would add another layer of complexity to determine the objects position relative to the hands. Second, our goal is to make this method work in real-time on a computer with a normal computing capability. The objects with simple, known shapes can be tracked without requiring heavy computations in contrast to hands with irregular shapes. The utility of the algorithm is demonstrated by the following four applications: the sport of Cup Stacking, the Soda Pop Coordination test, the Wechsler Block Design test, and a simple hand gesture test.

II. The Algorithm

A. Overview

Assessments of upper extremity motor skills often involves a set of objects that are manipulated by the person being evaluated or a sequence of tasks, such as extending the shoulder and twisting elbow [4], [5], [13], [36], [37]. Resulting measurements include the time for completion, accuracy, and extension/flexion range of each motion. Our approach aims to automate the evaluation process with real-time data collection by employing vision-based techniques using a standard webcam. The algorithm first identifies specific objects within a field of view, projects their position into 3D space based on known shape information, and then tracks them in real time. To simplify the processing time, the algorithm targets tracking objects being manipulated by or attached to human hands instead of directly tracking the hands. The output of this algorithm is the 3D pose of each object. The only requirements are that the item must have straight edges and it must be distinguishable from the rest of the environment by either its shape or color. Shape and color form a two-tiered classification structure that determines whether objects within a video frame are items of interest. These values can be altered depending on applications via calibration.

A major advantage of the presented algorithm over similar approaches [38], [39] is that it is versatile. The algorithm works with a variety of different markers without requiring reprogramming. The limiting factors are that the markers must be unique in the environment to avoid false detections. A detailed description of the algorithm is provided in the following subsections. Section II-B describes the item localization method based on the two-tiered classification scheme used to identify items of interest within the image and to detect the corner locations. Section II-C presents the pose estimation method using the corner points to project the item from the 2D image frame into the 3D real world frame using object shape information and internal camera properties. This process is called pose estimation and is a technique for extracting 3D information from a single camera frame. Lastly, Section II-D describes the camera calibration needed to introduce the algorithm to new markers and determine internal camera properties necessary for pose estimation. The codes were written in C++ and utilized OpenCV for computer vision implementations. The captured images are in the RGB format with a resolution of 640×480 pixels. Post-processing of the data for some applications was performed in MATLAB.

B. Item Localization

Localization is the process of identifying the location of the target item(s) in the image frame. We employ a two-tiered classification approach for localizing items of interest. A two-tiered system achieves a high degree of accuracy in identifying items as a result of the two different properties that are required to detect a matched item. Color and shape are used as the distinguishing properties. Color indicates the normalized color of the item within a certain color range. Shape is defined as the number and relative positions of an item's corner points. Explicitly, the algorithm searches images for items with a known normalized color, and then locates the edges of the items using morphological operations on the color regions. Those edges are then traversed to locate the items' corners. The resulting corners are the outputs of localization and can be used to estimate the 3D positions of the items using known shape information.

1) Color classification: The first step of localization is to segment the image in order to identify what parts of the input image could potentially be items of interest. Normalized color initially distinguishes the potential object regions within the image. It was chosen as the distinguishing property because it is not affected by adverse lighting conditions and represents the inherent color of an object [40]. Color normalization compensates for the intensity changes in lighting by forcing all intensity values to sum to 1. The well-known equations for normalizing the color at each pixel location in an image are used:

$$r = \frac{R}{R+G+B}; \quad g = \frac{G}{R+G+B}; \quad b = \frac{B}{R+G+B}$$

where $r + g + b = 1$. The intensity values correspond to the values of the three image planes (red (r), green (g), and blue (b)) that make up an image.

The color of an image is thresholded by examining each pixel's values to determine whether it falls within certain threshold ranges. A binary value of 1 or 0 is assigned based on whether it passes the threshold or not,

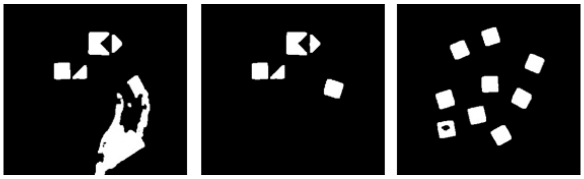


Fig. 1: Examples of binary images after color normalization.

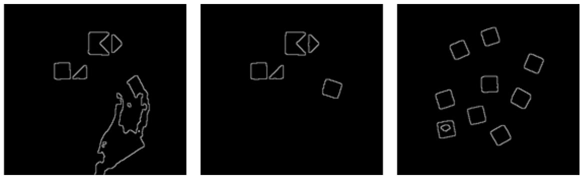


Fig. 2: Examples of edges found using the difference between a morphological dilation and the original image.

respectively. The ranges are defined by minimum and maximum values which are included in the objects color $(\{r_{\min}, r_{\max}\}, \{g_{\min}, g_{\max}\}, \{b_{\min}, b_{\max}\})$. These ranges can be easily identified for an object of interest by normalizing its color and finding the minimum and maximum color values for the object. Fig. 1 shows examples of the binary images resulting from color normalization.

The resulting binary image often requires additional processing to increase the accuracy. This process is necessary when color normalization fails to compensate for all imperfect lighting conditions or when the color threshold ranges are not completely accurate in reflecting the item's actual color. One technique is conditional dilation which can be beneficial when the color threshold detects only part of an item. It detects the rest of the item by expanding its area until it reaches the item's edges. The morphological operator of dilation is applied to the binary image but the results are only kept if the color values are close to the values of their neighboring pixels. Edges of objects are distinguishable by the dramatic change in the color range. Color values remain similar within the same item, but once dilation approaches an edge the values start to change quickly and exceed the acceptable range by the conditional dilation operator. In order to use this operator, the item's edges must be well defined.

2) Shape classification: The next stage of classification takes the outputs from color classification and further narrows down the regions of the image to detect the items of interest. Since color classification results in defined areas, the next step would often be blob detection. However, this is not the most convenient method in this case. Instead, the edges of the items are found and traversed in order to locate the corners of the item. The corners are then used to classify the item's shape as defined by the number of corners and their spacing. This method is chosen because it gives accurate positions of the corners used for classification that are essential for finding the object's 3D position. A morphological operator is used to find the object's edges through two steps. First, the color thresholded image is taken and a dilation operator is applied. The difference is then taken between the original image and the dilated image. A dilation operator expands the colored regions outward by operating on the image using a 3×3 rectangular structuring element. The result from taking the difference is an image containing only the edges of the colored areas. The edges are guaranteed to be 1-pixel thick, 4-connected, and form a closed loop. These three properties make them easy to traverse. The examples of these ideal edges are shown in Fig. 2.

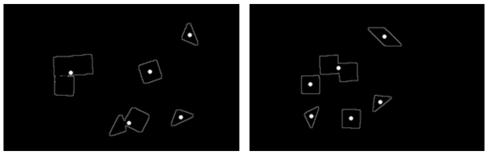


Fig. 3: Two images of blocks overlaid with centroids found after first traversal of each object's edges.

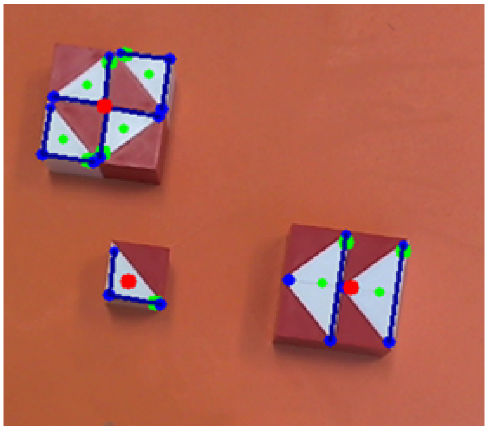


Fig. 4: A processed image showing the centroids of individual blobs and starting points (in green), overall centers (in red), and corners and block edges used to calculate tilt (in blue).

The identified edges are traversed with the aim of pinpointing the location of the corners (Fig. 3). It takes three traversals of an edge set to find these corners. On the first traversal, spurs are removed, the object's centroid is found, and the edges are put in a stack so they can be accessed easily on the second two traversals. The edges are traversed by first finding a point on the image that is part of an edge. The next point is found by checking the four coordinate directions in the order of up, right, down, and left and then moving in the first found direction that is part of the edge. Each point previously visited is added to a stack of edges so that it can be referenced later and the location in the image is blacked out so that it is no longer recognized as an edge. A spur is recognized to exist on the edge when a point, that is not the beginning, is found to have no neighbors. At that point, the path is retraced by popping values from the edge stack until a new point is found that has a neighbor, meaning that it originally had two neighbors. An edge is considered complete when it loops back to its starting location. The centroid of each region is calculated by keeping a running average of pixel locations.

After the first traversal of an object all the edge points are conveniently in a stack and the centroid has been calculated. The next traversal is used to find a point on the edge that is guaranteed to be a corner. Since the objects have straight edges, the edge location farthest from the object's centroid is guaranteed to be a corner point. This point is found by calculating the distance between the center and every point along the edge. The point that has the greatest calculated distance will be the corner point of interest. The last traversal finds all additional corners. They are found by moving along the edge and calculating the slope for each point edge. A constant slope designates the straight edge of an object and a rapid change in slope indicates a corner. After the corners are found they are amended by finding the intersection of lines fitted to the edges on either side of each corner. The resulting corner points are finally classified. If the number of points for an item is not equal to the expected number of corner points or the spacing of corners is not similar to that of the known shape, then the item can be conclude to not be an item of interest. For example, if the item of interest is a square then in order to be an object of interest there must be four evenly spaced corners. If it is found to be an object of interest then pose estimation can be used to get the object's 3D position. Fig. 4 shows an example of the processed image.

C. Position Estimation using Shape Information

The corners of the object, found through item localization, are used to estimate the position of the object in 3D space using known information about the object's shape and internal camera properties. The internal camera properties determine the perspective with which the camera views an object. By comparing the actual shape of an object with its warped shape within the camera frame, its pose relative to the camera can be determined. However, the relationship is nonlinear and typically cannot be solved directly. This can be circumvented using a variety of methods, including making additional assumptions about the object position or iterating to find the best values instead of solving directly. In an image frame, objects can be scaled and their perspective can be altered due to their relative position and orientation to the camera. If an assumption is made that the object's deformation in the image is either due to scaling or perspective then the equations can be simplified greatly. If it is assumed that the object has only been scaled, then the distance to the camera for all points on the object will be the same. At least two points are required to solve this system of equations, but the calculations can be made more accurate if more than two points are known.

The follow equation converts the $\{x, y, z\}$ image coordinate system to the $\{X, Y, Z\}$ real world coordinate system. The relationship can be described using their simple geometric relationship as shown in Fig. 5, given that the camera's focal length is broken down into f_x and f_y and the center of the image is at c_x and c_y. The Z axis is perpendicular to the image frame and both X and Y are parallel. The equations for this relationship are provided below and are rearranged so that they solve for the real world values. The two dimensions are independent and can thus be treated separately.

$$x = f_x \frac{X}{Z} + c_x \rightarrow X = (x - c_x)\frac{Z}{f_x}$$

$$y = f_y \frac{Y}{Z} + c_y \rightarrow Y = (y - c_y)\frac{Z}{f_y}. \tag{1}$$

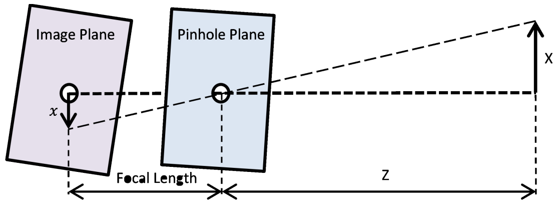


Fig. 5: The pinhole camera model in one dimension used for camera calibration.

There are three unknowns in (1), $\{X, Y, Z\}$. This problem is solved by using two points with a known relationship between each other and that are at the same distance from the camera. This provides a known distance between the points represented by the equation, $(\Delta X)^2 + (\Delta Y)^2 + (\Delta Z)^2 = d^2$ for a known d . The second simplification is that the points are at the same distance from the camera such that $Z_1 = Z_2 = Z$ since the Z-direction is perpendicular to the image. Under these assumptions, the real world coordinate is calculated by

$$X = (x - c_x)\frac{\sqrt{\frac{d^2}{\left(\frac{\Delta x}{f_x}\right)^2 + \left(\frac{\Delta y}{f_y}\right)^2}}}{f_x}$$

$$Y = (y - c_y)\frac{\sqrt{\frac{d^2}{\left(\frac{\Delta x}{f_x}\right)^2 + \left(\frac{\Delta y}{f_y}\right)^2}}}{f_y}$$

$$Z = \sqrt{\frac{d^2}{\left(\frac{\Delta x}{f_x}\right)^2 + \left(\frac{\Delta y}{f_y}\right)^2}} \tag{2}$$

where

$$\Delta X = (x_1 - x_2)\frac{Z}{f_x} = \Delta x \frac{Z}{f_x}$$

$$\Delta Y = (y_1 - y_2)\frac{Z}{f_y} = \Delta y \frac{Z}{f_y}$$

$$\Delta Z = Z_1 - Z_2 = 0.$$

D. Calibration

Calibration is an essential process to determine the conditions in which the camera is used and the properties of the item of interest. To locate an object, its color and shape must be known. The internal properties of the camera must also be quantified to determine how it views the item and to project the item from a 2D camera space into a 3D real space. The internal camera properties, or intrinsics, determine how a 3D object is projected into the 2D camera plane. The intrinsics includes the focal length and image center and are different for every camera. A camera can be represented by the pinhole camera model in which the light that the camera captures goes through a pinhole and is then projected onto the image plane. The focal lengths, f_x and f_y, are the distances in the x and y directions between the pinhole and the image center. The image center $\{c_x, c_y\}$ is the location of the pinhole projected onto the image frame. The geometric relationships between the 2D image and the 3D space are previously provided in (1). A commonly used object for determining the camera intrinsics is

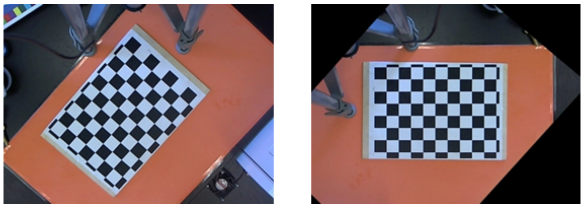


Fig. 6: A sample frame from a webcam showing the experimental set-up from a near-vertical camera view (left) and the transformed image to compensate for initial camera angle (right).

a checkerboard due to its defined number of points with known spacing. By analyzing the relative position of the checkerboard points within the image, the camera intrinsics can be found. Fig. 6 shows how camera intrinsics can be used to compensate for perspective and orientation undesirabilities. The left image frame shows the camera's perspective on the area and the right shows the perspective altered frame. It has been changed so that the corners of the checkerboard form a perfect square and aligned so that the work area lines up with the camera frame.

III. Algorithm Implementation

A. Overview

This section describes the applications of the computer vision algorithm previously described. The algorithm requires a uniquely colored square piece of paper to be placed visibly on the object to be tracked, or the object to contain a surface that is uniquely colored, and a camera to capture its motion. The paper/surface can be any color that is unique in the environment and the only requirement is that it stays visible throughout the motion. The versatility of the tracking algorithm is proven through its application to three different situations. The first is a sport played mostly by elementary school children called Cup Stacking. The second is a motor skill and coordination test developed by Hoeger & Hoeger called the Soda Pop Coordination test. The third is the Wechsler Intelligence Scale that is one of the most widely accepted psychological assessment tool. Among its subtests, we selected the Block Design test for the third application of our algorithm. For these applications, an automatic scoring system is implemented by identifying when certain events occur. In addition to these three specific examples, we also implemented the algorithm for potential applications in visual-motor integration assessment and gesture recognition.

B. Cup Stacking

Cup stacking (also called Sport Stacking) is an activity for individuals and teams in which specialized cups are used to create pyramids of three, six, and ten cups as quickly as possible. It is governed by the World Sport Stacking Association and a variety of studies have been conducted to assess its influence on motor skills [36], [41], [42]. Specifically, a study involving the second graders playing cup stacking for 15 minutes a day for 12 weeks showed that it might improve central processing and perceptual-motor integration skills [36]. Another study involving second and fourth graders playing cup stacking for 10-15 minutes a day for 3 weeks found no difference between a control group and a group participating in cup stacking [41]. It is also found that cup stacking is effective in improving hand-eye coordination and reaction time in second graders by playing the game for 20-30 minutes a day for 5 weeks [42]. It is notable that the sequences for stacking have a learning curve so cup stacking cannot be used to directly measure motor skills unless a training period is allowed.

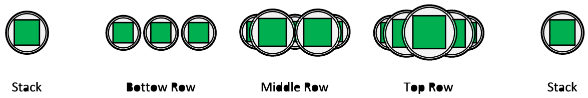


Fig. 7: Top view of the cups with green squares placed on the top illustrating five steps of six-cup stacking.

The scoring of cup stacking was automated by placing a marker on the top of each cup. The 3D position of each cup, (X, Y, Z), was found using pose estimation and then saved for further analysis. A six-cup stack game was employed as shown in Fig. 7. The automatic scoring was performed in real-time by recording when certain key actions occurred. The tasks included when the cups first started to move indicating the start of the activity, when three cups were placed as the base, when two cups were placed on top of the base, when the top cup was placed, and finally when all the cups come back together and stop moving indicating the activity is complete. A measure of the cup placement accuracy is determined by the straightness of the placement of cups in the bottom row and the relative angle between the bottom and middle rows, indicating how precisely the middle is placed relative to the bottom row. The automatic scoring component can easily be evaluated by comparing manually and automatically recorded trials. These values were compared for 91 different times and the resulting correlation is reasonable with an r-squared value of 0.9615 and an average error of 0.35 ± 0.27 seconds.

C. Soda Pop Coordination Test

The Soda Pop Coordination Test is a motor skills test that is a part of the American Alliance for Health, Physical Education, Recreation & Dance (AAHPERD) battery of tests. It is advantageous over similar tests because it uses commonly available materials and is easy to administer [37]. The test uses three soda pop cans and needs six marked locations on the table for the cans to be placed on, as shown in Fig. 8. In basic terms, the test involves flipping the soda cans over one at a time as fast as possible. Specifically, Can A is moved from position 1 to position 2, Can B is moved from position 3 is position 4, and Can C is moved from position 5 to position 6. Then the cans are moved back to their original positions in the reverse order. The hand must start with the thumb facing upward for the first set of movements and downward for the second set of movements. The test is usually scored using the time it takes to go back and forth twice and can be done for either the dominant or non-dominant hand.

The advantage of having an automated system for the Soda Pop Coordination test is that it increases the accuracy of scoring and makes data processing easier. Traditionally, the performance results would be a large amount of hand written

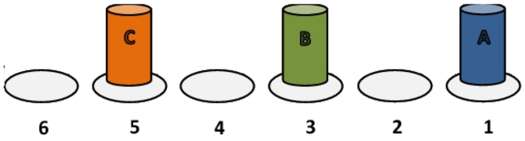


Fig. 8: Starting configuration for the soda pop coordination test with three soda cans and six locations

data (i.e. time, accuracy) that would have to be manually inputted into a computer. By having an automated system, the times are already saved on the computer and the mindless data entry step can be skipped, allowing for fewer opportunities for errors in recording. The Soda Pop Coordination test is usually administered before and after some training regimen to demonstrate how an action has improved a person's abilities [42], [43], [44]. It can also be administered to monitor coordination skills and then compared to or used to create standardized scores [37], [45]. Examples of before and after testing include a study to identify the effects of a 10 week Tai-Chi-Soft-Ball training on the physical functional health of Chinese adults [43], a 5 week study on second graders to identify the effect of 20-30 minutes of sport-stacking on hand-eye coordination and reaction time [42], and a study on the effect of a weight-bearing and water-based exercise program on osteopenic women [44]. Examples of using standardized scores include a study on the elderly, which showed the relationship between heart rate variability and coordination [45].

The test was automated by placing a marker on the top of Can A. The start time is set as the time the marker starts moving and the stop time is set as the time the marker comes back into view and stops moving. The marker will disappear as the cup is turned over and, in order to accommodate false starts, it is assumed to take more than two seconds to complete the test. Additionally, since the test has two sets of back and forth that count as one round, the numbers for the two consecutive times can simply be added together. The system was evaluated by comparing manually and automatically collected data to determine the accuracy and usability. Data was manually and automatically collected for 87 laboratory rounds of the Soda Pop Coordination test. The correlation between the manually and automatically collected data is 0.985 and the average difference in timing is 0.215 seconds.

D. Wechsler Block Design Test

The Wechsler Intelligence Scale for Children (WISC) and the Wechsler Adult Intelligence Scale (WAIS) are widely accepted psychological assessment tests used to measure intelligence in children and adults that were initially developed by David Wechsler in the 1930s [46], [6], [7]. Both scales contain a subtest called the Block Design test that measures a person's non-verbal conceptualization, spatial visualization, and fine-motor control [47]. The Block Design test was first proposed by Kohs in 1923 [48], but has been incorporated in some form into most intelligence tests. The WISC and WAIS subtests themselves involve recreating 2D red-and-white geometric patterns using 3D cubes that have red, white, and red-and-white sides. The patterns can be made up of two, four, or nine blocks and a score is awarded for each pattern based on the time taken to complete the assembly and whether the final assembly is correct [6], [7]. Typically when this test is administered, a trained professional must be present to walk the testee through the process by keeping track of completion times, recording incorrect answers, scoring the test, and monitoring the test taker for any psychological clues.

In this test, the algorithm was implemented to directly track the blocks instead of placing separate markers on them because the blocks themselves satisfy the requirements for serving as a marker. The blocks have sides that appear as triangles and squares when either white or red color is tracked, as well as being able to form more complex shapes by putting the blocks together. Scoring requires additional considerations because the system must recognize whether a testee has successfully created a pattern using multiple blocks. This means that the resulting position of the blocks must be used to estimate the pattern created by the blocks. The scoring process involves overlaying a grid over the found blocks and determining the color layout within the grid to match to the pattern. The start time of a trial was indicated by the blocks being dispersed throughout the environment and is marked as complete when the blocks form the goal pattern of that trial. If a pattern is not completed successfully, then the time is stopped and marked incomplete when the blocks are dispersed for the next pattern. The score is assigned by the same conventions as in the WISC and WAIS block design tests. The automation was tested by comparing manually and automatically scored tests showing 100% accuracy.

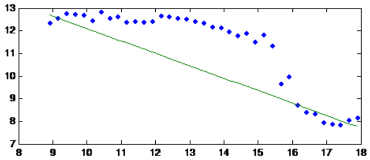
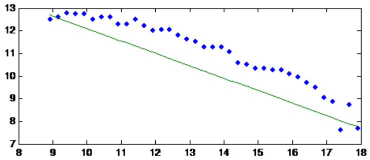
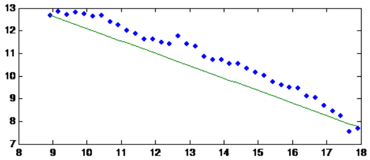
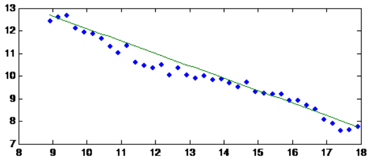
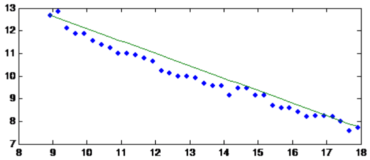
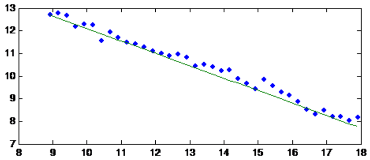
E. Visual-Motor Integration Test

A part of motor skills is reflected by how well a person can trace lines and shapes in 2D. The closeness of a followed path to the ideal path and steadiness of the movements reflect the motors skills of the person in terms of how advanced they are in their motor development or if they have any difficulties with any of their individual joints or muscles. The idea is similar to that of the Beery-Buktenica Developmental Test of Visual-Motor Integration (Beery VMI) where the subject must copy or trace lines and shapes using a pencil [49]. For our demonstration, a cup is used instead of a pencil to trace out a pattern on the table and shapes in the air. The path could be anything as long as its shape is known so that an ideal path is available for comparison. Table I shows six trails of drawing a straight line between two points using a cup as a marker. A correlation between the actual position data and an ideal fitted straight line was analyzed by performing a linear regression between the two. A higher value of r indicates the movement trajectory was closer to the given straight line.

F. Gesture Recognition

Gesture recognition aims to classify the motion that a person is performing [49], [50]. It has a wide range of applications including aids for the hearing impaired, interpreting sign language, lie/stress/emotional state detection, and controls or tools for interaction with virtual environments [49]. A variety of methods can be used to interpret gestures including principal component analysis, the CONditional DENSity PropagATION (CONDENSATION) algorithm, Kalman filtering and more advanced particle filtering, and hidden Markov models [49]. The goal of this application is to create a simple gesture

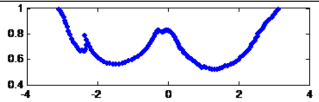
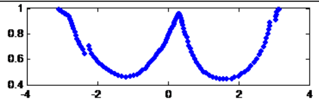
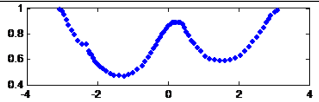
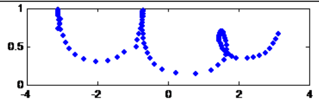
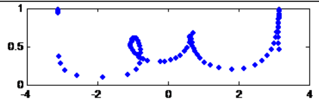
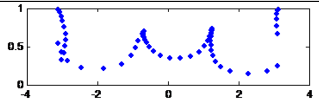
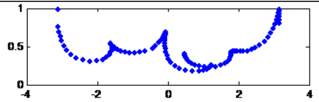
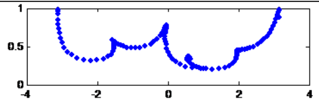
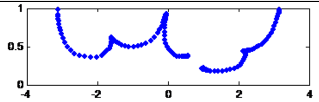
TABLE I: Paths exhibiting a range of different accuracies between two points shown in graphs of points and straight ideal lines along with the calculated correlation values for the match between the two.

Correlation coefficient	Movement trajectory
$r = 0.8539$	
$r = 0.9562$	
$r = 0.9827$	
$r = 0.9930$	
$r = 0.9889$	
$r = 0.9930$	

recognition tool that can identify the motion of tracing the geometry of a shape. It is also desired that it is not affected by the speed or the size of the motion but is simply unique to the shape or form of the motion.

Only a couple of distinct shapes were explored for this section, so a simple method was chosen for recognition. The motions were to draw a circle, triangle, and square in the air and the method used to recognize the shapes was a shape descriptor technique called shape signatures [51]. Shape signatures represent an object's shape as a one dimensional function of its edge points. A variety of different methods can be used to create this function but a common method, which is used here, is the distance of the boundary points and angle relative to their centroid. The signature is made scale invariant by dividing all distances by the maximum distance and is made orientation invariant by finding the angular position of the maximum point and making the function start at this value.

TABLE II: Signatures for the three shapes (circle, square, and triangle) and calculated feature values and logic gate outputs using the threshold values of $P_1 = 0.4$ and $P_2 = 0.6$.

Motion Trajectory	P_1	P_2	Output
	0.5237	0.8350	*Circle*
	0.4501	0.9350	*Circle*
	0.4740	0.8977	*Circle*
	0.1545	0.2119	*Triangle*
	0.1047	0.3606	*Triangle*
	0.1569	0.3932	*Triangle*
	0.1825	0.6952	*Square*
	0.2155	0.7871	*Square*
	0.1849	0.9363	*Square*

The signature can then be analyzed to find the number of corners mapped to a function. In this case, the signature is simply examined at key locations that distinguish the different shapes. The first key feature is the relationship between the minimum and maximum value of the radius (R_{min}, R_{max}). This distinguishes circles (or in this case ellipses) from the squares and triangles. Unless the eccentricity is high, the ratio will be significantly higher for circles then for the other two. The second feature is the behavior of the shape at an angle of zero. Circles have extreme points at angles of π and 0, so the behavior at 0 should be high. Squares have extreme points at π, $\pi/2$, 0, and $\pi/2$, so the behavior at 0 should be high. Finally, triangles have extreme points at π, $2\pi/3$, and $-2\pi/3$, so the behavior at 0 should be low. For these three shapes, two features effectively take care of all possible cases. If more shapes need to be identified then additional features would become necessary, but would be easy to add to the current framework. Table II shows recognition results for each shape. P_1 and P_2 are calculated by

$$P_1 = \frac{R_{min}}{R_{max}}; \quad P_2 = \frac{R_{center}}{R_{max}},$$

and the shape is recognized as a *circle* if $P_1 > 0.4$ and $P_2 > 0.6$, a *triangle* if $P_1 < 0.4$ and $P_2 < 0.6$, or a *square* if $P_1 < 0.4$ and $P_2 > 0.6$.

IV. Conclusion and Discussion

This paper presented an integrated low-cost, real-time vision processing algorithm that can be used for a variety of assessment tests for upper extremity motor skills that involve object manipulation. While individual layers of the algorithm utilize existing techniques, the main contribution of this paper lies in the proper integration of these techniques keeping the computational cost low for target clinical and educational applications. The algorithm was implemented in four well-known games/tests and a simple gesture recognition application for demonstrating its potential utility. When such motor assessment tests need to be periodically administered to an individual or to a large group of people, automating the entire process can significantly reduce the time, cost, and labor intensity while also improving the quantity and quality of the measurable data. The specific applications presented in this paper were carefully selected to cover a broad range of motor skill assessment tests so that one can easily take it into use.

The presented algorithm requires comparison with other vision-based object tracking algorithms to prove its time efficiency. To further improve the versatility of the algorithm, another layer of prior image processing can be added for automatically determining the color threshold range instead of using a pre-defined value so that any arbitrary objects can be detected and tracked as long as they are distinguishable from the environment. In addition, benefits expected by the algorithm implementation needs to be verified through human subject studies involving non-technical administrators (e.g. teachers, parents, and clinicians) and potential testees (e.g. students, children with varying cognitive/motor skills, and older adults). Our ongoing work involves human subject evaluation and cost analysis in addition to continuous improvements in the algorithm.

References

[1] G. H. Noritz, and N. A. Murphy, “Motor delays: early identification and evaluation,” Pediatrics, 131(6): e2016-e2027, 2013.

[2] American Academy of Pediatrics, Committee on Children with Disabilities, “Developmental surveillance and screening of infants and young children,” Pediatrics, 108: 192-196, 2001.

[3] Division of Health Policy Research, “Identification of children ¡ 36 months at risk for developmental problems and referral to early identification programs.,” American Academy of Pediatrics, Periodic Survey of Fellows., 2003.

[4] Beery-Buktenica Developmental Test of Visual-Motor Integration, 6th Edition (BEERY VMI),” Pearson Education, Inc.

[5] C. DeMatteo, M. Law, D. Russell, N. Pollock, P. Rosenbaum and S. Walter, “QUEST: Quality of Upper Extremity Skills Test,” Canchild Centre for Childhood Disabilities Research, Hamilton, Ontario, Canada, 1992.

[6] D. Wechsler, “Manual for Wechsler Adult Intelligence Scale - Revised”, The Psychological Corporation, New York, 1981.

[7] D. Wechsler, “Wechsler Intelligence Scale for Children - Revised”, The Psychological Corporation, New York, 1971.

[8] L. First and J. Palfrey, “The infant or young child with developmental delay,” The New England Journal of Medicine, 330(7): 478-483, 1994.

[9] F. P. Glascoe, M. Foster and W. Wolraich, “An economic analysis of developmental detection methods.,” Pediatrics, 99: 830-837, 1997.

[10] C. A. Brogan , “The pathway to care for children with autism spectrum disorders aged 0 to 12 years.,” Glasgow: National Autistic Society Scotland, p. 2001.

[11] T. H. Lee, C. M. Blasey and J. Dyer-Friedman, “From research to practice: Teacher and pediatrician awareness of phenotypic traits in neurogenetic syndromes,” American Journal on Mental Retardation, vol. 10, no. 2, pp. 100-106, 2005.

[12] R. Parasuraman and V. Riley, “Humans and Automation: Use, misuse, disuse, abuse,” Human Factors, vol. 39, no. 2, pp. 230-253, 1997.

[13] V. Mathiowetz, G. Volland and K. Weber, “Adult Norms for the Box and Block Test of Manual Dexterity,” American Journal of Occupational Therapy, vol. 39, pp. 386-391, 1985.

[14] J. Desrosiers and G. Bravo, “Validation of the Box and Block Test as a measure of dexterity of elderly people: reliability, validity, and norms studies,” Arch Phys Med Rehabil, vol. 75, pp. 751-755, 1994.

[15] C. Hekimian-Williams, “Accurate Localization of RFID Tags Using Phase Difference,” in Proceedings of the IEEE International Conference on RFID, Orlando, FL, 14-16 April 2010.

[16] A. W. Reza and T. K. Geok, “Objects tracking in a dense reader environment utilizing grids of RFID antenna positioning,” International Journal of Electronics, vol. 96, no. 12, pp. 1281-1307, 2009.

[17] D. Jeong, E. Kerci, and L. Lee, “ TaG-Games: Tangible Geometric Games for Assessing Cognitive Problem-Solving Skills and Fine Motor Proficiency,” in Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligence Systems, pp. 32-37, 2010.

[18] L. Buechley and M. Eisenberg, “Boda Blocks: A Collaborative Tool for Exploring Tangible Three- Dimensional Cellular Automata,” in Proceedings of the 8th International Conference on Computer Supported Collaborative Learning, pp. 102-104, 2007.

[19] R. Watanabe, Y. Itoh, M. Asai, Y. Kitamura, F. Kishiro, and H. Kikuchi, “The Soul of ActiveBlock: Implementing a Flexible, Multimodal, Three-Dimensional Spatial Tangible Interface,” in Proceedings of the ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, 2004.

[20] D. Anderson, J. L. Frankel, J. Mark, D. Leigh, K. Ryall, E. Sullivan and J. Yedidia, “Building Virtual Structures with Physical Blocks,” in Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology, 1999.

[21] L. Terrenghi, M. Kranz, P. Holleis, and A. Schmidt, “A cube to learn: a tangible user interface for the design of a learning appliance,” Pers Ubiquit Comput, (10): 153-158, 2006.

[22] K. Camarata, E. Y. Do, M. D. Gross, and B. R. Johnson, “Navigational Blocks: tangible navigation of digital information,” in Proceedings of International Conference on Computer Human Interaction (CHI), 2002.

[23] M. G. Gorbet, M. Orth, and H. Ishii, “Triangles: Tangible Interface for Manipulation and Exploration of Digital Information Topograph,” in Proceedings of CHI, pp. 18-23, 1998.

[24] B. Schiettecatte, and J. Vanderdonckt, “ AudioBlocks: a Distributed Block Tangible Interface based on Interaction Range for Sound Design,” in Proceedings of the 2nd International Conference on Tangible and Embedded Interaction (TEI), 2008.

[25] H. Newton-Dunn, H. Nakano, and J. Gibson, “Block Jam: A Tangible Interface for Interactive Music,” in Proceedings of the Conference on New Interfaces for Musical Expression, pp. 170-177, 2003.

[26] S. Alers, E. I. Barakova, “Multi-Agent Platform for Development of Educational Games for Children with Autism,” in Proceedings of Games Innovations Conference, Intl. IEEE Consumer Electronics Society, 2009.

[27] J. P. Wachs, M. Kolsch, H. Stern, and Y. Edan, “Vision-based Hand-gesture Applications,” Communications of the ACM, 54(2): 60-71, 2011.

[28] H. Kato, M. Billinghurst, I. Poupyrev, K. Imamoto, and K. Tachibana, “Virtual Object Manipulation on a Tabletop AR Environment,” in Proceedings of IEEE and ACM International Symposium on Augmented Reality, 2010.

[29] A. D. Wilson, “Depth-Sensing Video Cameras for 3D Tangible Tabletop Interaction,” in Proceedings of IEEE International Workshop on Horizontal Interactive Human-Computer Systems, Newport, RI, 2007.

[30] A. D. Wilson, “PlayAnywhere: A Compact Interactive Tabletop Projection-Vision System” Proceedings of the 18th annual ACM symposium on User interface software and technology, 2005.

[31] A.D. Cheok, X. Yang, Z. Z. Ying, M. Billinghurst, and H. Kato, “Touch-space: Mixed reality game space based on ubiquitous, tangible, and social computing,” In Proceedings of the 2004 ACM SIGCHI International conference on Advances in Computer Entertainment Technology, pp. 117-126, 2003.

[32] B. H. Thomas, “A survey of visual, mixed, and augmented reality gaming,” Computers in Entertainment Magazine, 10(3), Article 3, 2012.

[33] D. Jeong, E. Kerci and K. Lee, “TaG-Games: tangible geometric games for assessing cognitive problem-solving skills and fine motor proficiency,” IEEE MFI, pp. 32-37, 2010.

[34] B. Floyd, D. Jeong and K. Lee, “Geometric Games for Assessing Cognitive, Working Memory, and Motor Control Skills,” in Tangible, Embedded, and Embodied Interaction, Kingston, ON Canada, 2012.

[35] R. Y. Wang and J. Popovic, “Real-Time Hand-Tracking with a Color Glove,” ACM Transactions on Graphics, vol. 28, no. 3, 2009.

[36] Y. Li, D. Coleman, M. Ransdell, L. Coleman and C. Irwin, “Sport Stacking Activities in School Children’s Motor Skills Development,” Perceptual and Motor Skills, vol. 113, no. 2, pp. 431-438, 2011.

[37] E. Hoeger and S. Hoeger, Principles and Labs for Fitness and Wellness, Belmont, CA: Wadsworth/Thomson Learning, 2004.

[38] A. Yilmaz, O. Javed and M. Shah, “Object tracking: A survey,” ACM Comput. Surv., vol. 38, no. 4, 2006.

[39] W. Hu, T. N. Tan, L. Wang and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” IEEE Transaction on Systems, Man, and Cybernetics Part C - Applications and Reviews, vol. 34, no. 3, pp. 334-352, 2004.

[40] M. J. Swain and D. H. Ballard, ”Color indexing,” Internal Journal of Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.

[41] M. Hart, L. Smith and A. DeChant, “Influence of Participation in a Cup-Stacking Unit on Timing Tasks,” Perceptual and Motor Skills, vol. 101, pp. 869-876, 2005.

[42] B. Edermann, S. Murray, J. Mayer and K. Sagendorf, “Influence of Cup Stacking on Hand-Eye Coordination and Reaction Timeof Second-Grade Students,” Perceptual and Motor Skills, vol. 98, pp. 409-411, 2004.

[43] M. Lam, S. Cheung and B. Chow, “The effects of Tai-Chi-Soft-Ball training on physical functional health of Chinese older adult,” J. Hum. Sport Exerc, vol. 6, no. 3, 2011.

[44] G. Bravo, P. Gauthier, P. Roy, H. Payette and P. Gaulin, “A Weight-Bearing, Water-Based Exercise Program for Osteopenic Women: Its Impact on Bone, Functional Fitness, and Well-Being,” Arch Phys Med Rehabil, vol. 78, no. 12, pp. 1375-80, 1997.

[45] R. H. Wood, J. M. Hondzinski and C. M. Lee, “Evidence of an association among age-related changes in physical, psychomotor and autonomic function,” Age and Ageing, vol. 32, no. 4, pp. 415-421, 2003.

[46] G. Frank, “The Wechsler Enterprise: An Assessment of the Development, Structure, and Use of the Wechsler Tests of Intelligence”, Pergamon, Oxford, 1983.

[47] J. Sattler, “Assessment of childern’s intelligence”, Saunders, Philadelphia, 1974.

[48] S. Kohs, “Intelligence Measure”, Macmillan, New York, 1923.

[49] S. Mitra and T. Acharya, “Gesture Recognition: A Survey,” IEEE Transaction on Systems, Man, and Cybernetics Part C - Applications and Reviews, vol. 37, no. 3, pp. 311-324, 2007.

[50] J. Daugman, “Face and Gesture Recognition: Overview,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 675-676, 1997.

[51] S. Belongie, J. Malik and J. Puzich, “Matching Shapes,” in 8th IEEE International Conference on Computer Vision, Vancouver, Canada, 2001.

Maximizing Throughput of SW ARQ with Network Coding through Forward Error Correction

Farouq M. Aliyu, Yahya Osais, Ismail Keshta, Adel Binajjaj
Department of Computer Engineering
King Fahd University of Petroleum and Mineral Resources
Dhahran, Saudi Arabia

***Abstract*—Over the years, several techniques for improving throughput of wireless communication have been developed in order to cater for the ever increasing demand of high speed network service. However, these techniques can only give little improvement in performance because packets have to be delivered as is. As such researchers have begun thinking outside the box by proposing ideas that require relay nodes to temper packets' contents in order to improve the throughput of a network. One of the state of the art techniques in this field is called Network Coding (NC). NC is a state of the art technique that allows relay nodes linearly combine two or more packets in a way they can be recovered upon reaching their destination. However, increasing packet size increases possibility of error affecting it. In this paper, the authors decide to investigate whether adding data recovery technique can improve the performance of a network that uses network coding, if it can, by how much can it? Is it worth the trouble? In order to answer these questions, the authors carried out a quantitative analysis of throughput in a Stop-and-Wait Automatic Repeat reQuest (SW-ARQ) data transmission system with Network Coding (NC) and Forward Error Correction (FEC). Vandermonde matrix is chosen as the coding technique for this research because it has both NC and data recovery characteristics. Python programming language is used to develop three Discrete Event Simulations: SW-ARQ without any NC, SW-ARQ with NC and SW-ARQ with NC and FEC. The obtained results show that SW-ARQ with NC and FEC is superior to traditional SW-ARQ in terms of throughput, especially in channels with high error rates.**

***Keywords*—Network Coding; Automatic repeat request (ARQ); Stop-and-Wait (SW); Vandermonde Matrix**

I. Introduction

Non-ideal behavior of communication channel causes received data to sometimes change from its original form, thus leading to misinterpretation. Automatic Repeat reQuest (ARQ) is one of the basic error control protocols used to provide reliable communication between two wireless devices. There are three main ARQ systems namely; Go-back-N, Selective Repeat and Stop-and-wait [1].

In stop-and-wait (SW) ARQ, transmitter sends a frame and then waits until it receives a reply (i.e. Acknowledgment (ACK) or Negative ACK (NACK)) for the transmitted packet from the receiver. Although, SW-ARQ is simple to implement and guarantees that packets are received in order, it is not efficient because of the wasted time during waiting for replies. Thus, it has low throughput [2]. Throughput can be defined as the average rate of successful data transmission over a network, and it is normally given in bits per second (bps). Go-back-N ARQ and selective repeat ARQ protocols manage to reduce ACK overhead by transmitting multiple frames without waiting for their ACK. For further information on ARQ protocol refer to [1] [2].

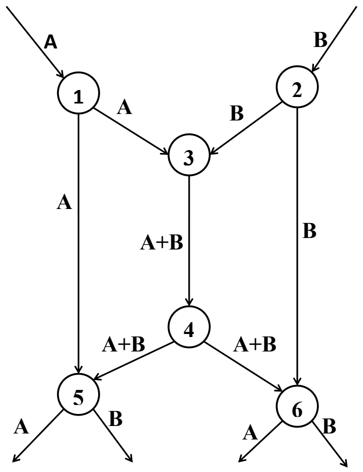


Fig. 1: Butterfly Network Coding in a multicast network

Other data transmission techniques for improving overall network's throughput are Network Coding (NC) and Forward Error Correction (FEC). NC was invented by Ahlswede *et al* [3] in 2000. It is an in-network data processing technique that allows intermediate nodes to aggregate two or more packets into one before forwarding it [4]. Figure 1 shows a butterfly wireless network with the vertices representing routing nodes and the edges representing the path of the packets. NC has several advantages; studies have shown that NC enhances the overall network throughput [1][5], it increases performance in multi-rate networks [4] and increases robustness of the network [6].

Forward Error Correction (FEC) on the other hand, is an error control technique where redundant data is systematically generated and embedded in a packet such that it can be regenerated in the event of error during transmission [7]. In FEC, the transmitter takes N data symbols and encode them with M parity symbols to form $N + M$ new symbols before transmitting them. At the receiver the original data symbols can be reconstructed as long as N out of the $N + M$ received symbols are error free. For more information on FEC refer to [7].

This paper extends the work in [8], where the throughput of a SW-ARQ with NC was investigated. Here we investigate the effects of NC when combined with FEC on the throughput of SW-ARQ using Python high level programming language. The remaining part of this paper is as follows: Section II reviews some of the related work in ARQ and NC. Section III provides detailed information on the developed network model and compares it with the work in [8] where necessary. In Section V analysis of the results obtained and comparison with results in [8] is carried out. Finally, in Section VI conclusion(s) was/were drawn based on the results that were obtained.

II. Related Work

De Vuyst al. in [9] present an analysis of the SW-ARQ protocol, in which it was pointed out that errors occur in bursts as packets are transmitted from the transmitter to the receiver and that the probability of receiving an erroneous packet depends on state of the channel when the packet was transmitted. Gilbert *et al* [10] divide these states into two; GOOD state and BAD state — using two-state Markov Chain to model the channel. Their result shows that the delay (which is inversely proportional to throughput) increases sharply with increase in the capacity of the channel. In order to compensate for the drop in throughput due to increase in capacity, techniques like Forward Error Correction(FEC) and Network Coding (NC) are often used.

In [11], the authors investigate the effect of network coding (NC) on throughput of the three basic ARQ systems. Their findings show that NC significantly improves the throughput of all the three ARQ protocols. Furthermore, studies show that wireless networks using NC give better throughput even though the complexity of the system increases. A novel Automatic Repeat reQuest (ARQ) system for cooperative wireless networks is introduced by Antonopoulos *et al* [12] in 2011, where cooperative and network coding techniques are combined in order to enhance the system's performance. They are able to obtain 85% bandwidth improvement due to the reduction of the total number of transmissions. Li *et al* [13] propose system for wireless broadcast system based on the random network coding using two-State Markov Chain to model the channel. The authors analyze the throughput of a typical SR-ARQ using; Linear Network Coding (LNC) and Random Network Coding (RNC) and it is found that random network coding gives better throughput, especially in the case of a system with a large number of receivers. Liu *et al* [14], introduce NC-ARQ system based on Two-State Markov channel for Cognitive Radio (CR) broadcast, where lost packets are XORed by CR base-station forming a new packet. The new packet is then broadcasted by CR base-station to all of CR users. The results show noticeable improvement in throughput, especially when there are large numbers of CR users. On the contrary, authors in [15] study and analyze the steady-state throughput of SW-ARQ with NC using finite state machine, results obtained show that as the number of incoming links to the base-station increase a bottleneck in information delivery is formed.

Algorithm 1: Pseudo-code for SW ARQ-NC used in protect [8]

```
$pktCount \leftarrow 0$
clear buffer
while ($pktCount < n$) do
    wait for packet
    buffer ← packet
    $pktCount \leftarrow pktCount + 1$
end while
Matrix ← generated Vodermonde Matrix(buffer)
$i \leftarrow 0$
while ($i \leq pktCount$) do
    reply ← transmit(Matrix[$i$])
end while
```

Alsebae *et al* [8], study the effect of network coding (NC) system on the throughput of SW-ARQ. The system is simulated using MATLAB SimEvents Discrete Event simulation toolbox. The system is described in Algorithm 1. It measures the throughput of two nodes with similar function to node 3 and 4 in Figure 1. In their work, the transmitter waits for n packets, which are then converted into an $n \times n$ Vandermonde matrix (see Equation (1)). Each row (or block) of the matrix is considered as a new packet. These new packets are then transmitted to the receiver where the original n packets are regenerated. The researchers conclude that SW-ARQ with NC has better throughput, particularly in cases where the channel has high error rate. Finally, they postulated that it could give higher throughput than that which they have accomplished. This led to the research reported in this paper, where FEC abilities of Vandermonde matrix have been exploited. This approach is inspired by the fact that probability of error increases exponentially with increase in packet size. Therefore, there is need to add FEC to the protocol in order to increase its throughput as we shall see in Section III.

III. SW-NC with Forward Error Correction

$$V_{(nxn)} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \dots & \alpha_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \dots & \alpha_n^{n-1} \end{pmatrix} \tag{1}$$

The proposed system is simulated using two nodes, node 3 and 4 of Figure 1. The system has a transmitter which collects n packets. It converts the n packets into a $n \times n$ Vandermonde Matrix. The matrix takes the general form shown in Equation 1, where each row is a block of data that can be transmitted as a packet. At the receiver, the blocks are broken down into their original components (i.e. the earlier n packets).

Algorithm 2 provides a general overview of the proposed SW-ARQ with NC and FEC. The system consists of three main simulation components namely;

A. Transmitter Model

This part is modeled to contain two sub-modules namely *packet generation* and *encoding*. Packet generator, generates packets at a fixed rate λkbps. Once the time (i.e. $\frac{packet\_size}{\lambda}$) has pass, the packet generation sub-module will add a new data packet to the transmitter's buffer. Every data packet is saved in the format: [seqnum, payload], where seqnum is the packets sequence number and payload is the size of the packet in

Algorithm 2: Pseudo-code for proposed system

```
$pktCount \leftarrow 0$
clear buffer
while $(pktCount < n)$ do
    wait for packet
    buffer ← packet
    $pktCount \leftarrow pktCount + 1$
end while
Matrix ← generated Vodermonde Matrix(buffer)
$i \leftarrow 1$
while $(i \leq pktCount)$ do
    reply ← transmit(Matrix[$i$])
    if (reply = NACK and $i \neq pktCount$) then
        $i \leftarrow i + 1$
    else
        if (reply = ACK) then
            exitWhile
        end if
    end if
end while
```

bits. This process is repeated until the total number of packets generated equals to a certain number of the pre-programmed packets (n).

After the required numbers of packets were generated, the transmitter forwards them to the Vandermonde Matrix Encoder. There, the packets are stripped off of their headers and trailers before they are converted into blocks representing the rows of the Vandermonde matrix as shown in Equation 3. The Vandermonde Matrix is formed purely from the packets payload. The advantage of Vandermonde Matrix is that the encoded data packets are linearly independent. Hence the receiver is able to recover the packets. The mathematical equation representing how packets are converted to Vandermonde Matrix is as follows:

let $P_1, P_2, \ldots P_n$ be packets generated and $b_1, b_2, \ldots b_n$ be the blocks generated then,

$$\begin{aligned} b_1 &= P_1 + P_2 + \cdots + P_n \\ b_2 &= P_1^2 + P_2^2 + \cdots + P_n^2 \\ &\vdots \\ b_{n-1} &= P_1^{n-1} + P_2^{n-1} + \cdots + P_n^{n-1} \end{aligned} \tag{2}$$

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ P_1 & P_2 & \ldots & P_n \\ P_1^2 & P_2^2 & \ldots & P_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ P_1^{n-1} & P_2^{n-1} & \ldots & P_n^{n-1} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{3}$$

Note that the first row of the Vandermode Matrix is always filled with ones as such this row is never transmitted. On the contrary, it is transmitted in [8]. This can be seen as an overhead since this information is redundant.

After each transmission, the transmitter pauses until it receives the ACK or NACK from the receiver. On one hand, if the transmitter receives NACK, it transmits the next block in the hope that a different column might be in error, thereby allowing the receiver to apply forward error correction. On the other hand, if the block in question is the last block, then the transmitter keeps sending it (since there are no more blocks to send) until it receives an ACK. As a rule of thumb, transmission is over whenever ACK is received, because it signifies that all packets can be recovered from the blocks received. However, the authors in [8] fail to take the advantage of this unique characteristic of the Vandermonde matrix. Thus, all columns are transmitted.

B. Channel Model

In order to simulate corrupt frames accurately, the channel is modeled using *Binary Symmetric Channel* (BSC) [16][17]. BSC is an independent and identically distributed (i.i.d.) channel with the probability of a given bit "*flipping*" as ϵ also known as *Bit Error Rate* (BER). Therefore, the probability of finding an error in a given frame can be represented by Equation 4.

The channel module keeps checking for the presence of data. Once data is placed on the channel, the channel checks it first; if data is an ACK from the receiver it is passed to the transmitter directly without going through the error simulation module, because ACK/NACK packets are so small in size that error has negligible effect on them as shown by Equation 4. However, if the packet is not an ACK/NACK, the channel passes it to the error sub-module, where *Monte Carlo* method is applied to it in order to randomly choose the blocks to be in error based on the frame error rate equation (i.e. Equation (4)). Frame Error Rate (FER) is the probability that one or more bits in a frame are in error.

$$FER = (1 - (1 - BER)^k) \tag{4}$$

Where, FER = Frame Error Rate
BER = Bit Error Rate
k = Number of bits in a frame

C. Receiver Model

Finally, the packet reaches the receiver which hands it to a sub-module called the *Error Checker*, where the received packet is checked for errors; if error(s) is/are found, then the columns of the Vandermonde Matrix collected so far are checked. If all elements of a column are in error (as in Equation 6) then the original packets cannot be recovered. The receiver is notified and it sends a NACK packet to the transmitter, but saves the corrupt packet in the hope that it can help in future error corrections. Corrupt packets are discarded and new packets are requested in the system proposed by [8].

$$V_{(3x3)} = \begin{pmatrix} 1 & 1 & 1 \\ Error1 & 6 & 3 \\ 4 & 36 & Error2 \end{pmatrix} \tag{5}$$

$$V_{(3x3)} = \begin{pmatrix} 1 & 1 & 1 \\ Error1 & 6 & 3 \\ Error2 & 4 & 36 \end{pmatrix} \tag{6}$$

In a nutshell, receiver attempts forward error correction on the Vandermonde Matrix, if that fails the transmitter then sends the next block of data. However, if it is the last block the transmitter keeps sending it until an ACK is received. Equation 5 and 6 have illustrate scenarios where original packets can and cannot be recovered respectively.

D. *FEC technique*

Vandermode matrix can be used in conjunction with other error detection techniques to develop a packet recovery system [18]. As shown in Equation 2 and 3, packets (P) from other nodes are encoded into Vandermonde matrix. Each row (also known as Block (b)) is transmitted across the network as a packet. At the receiver these blocks are checked for errors before the original packets are finally recovered.

Suppose an element in the x row and y column of the Vandermonde matrix is represented by $\alpha_{x,y}$, then the nth packet encoded can be retried using Equation 9.

$$P_n = \alpha_{2,n} \tag{7}$$

$$\alpha_{2,n} = \sqrt[x-1]{\alpha_{x,n}} \quad where \quad x > 1 \tag{8}$$

$$\Rightarrow P_n = \sqrt[x-1]{\alpha_{x,n}} \quad where \quad x > 1 \tag{9}$$

IV. Simulation

Although Matlab Simevents was used in [8], Python programming language was used in this research. It allows the programmer limitless flexibility and levels of conception. However, in order to ensure accuracy the system proposed by [8] was first reproduced and its codes was later tweaked to develop our proposed system.

Discrete Event Simulation (DES) approach was used [19][20]. The events used are described in Section III. Three separate codes were written base on the network setup shown earlier in Figure 1. The first code simulates the network without any network coding technique added. The second simulates the network with Vandermonde matrix used as means of coding the network. While the third code uses the Vandermonde matrix as both network-coding and data recovery technique. Each of the three codes were then simulated to thirty seconds of simulation time. The results where then printed in the form of graphs that are presented in Section V, this was done with the help of the ”matplotlib” python library [21]. The variables were also exported using the programs IDE and further analysis was carried out on the data. This is possible because Python(x,y) IDE was used for the simulation [22].

Table I enlists parameters used in the simulation of the SW-ARQ communication system with and without NC. These parameters are exact replica of those used in [8], which allow us to compare the performance of the two simulations. However, it is worth noting that packet generation rate (λ) is changed from 50 to $100 packets/s$ in order to ensure maximum performance for all three networks as indicated by Figure 7.

V. Discussion

In this section the throughput performance of the proposed system is presented and analyzed. The section also draws out some possible applications of the proposed system.

TABLE I: Parameters used in simulation

Parameter	Value
pkt_size	1000 bits
lamda (λ)	100 Packets/s (packet generation rate)
Rate	10 Mbps (The system bit rate)
Tprop	15 ms
Tsim	40000 ms (Simulation time)
FER	Forward error rate (values used: 0, 0.1, 0.6, 0.9)
ack_waiting_time	default setting, 0
n	5 blocks per code

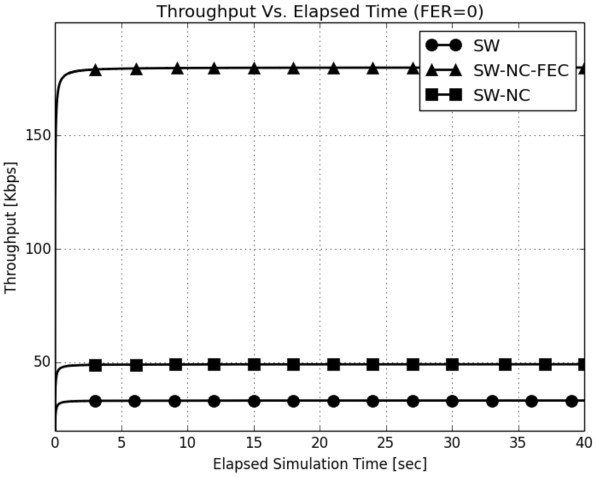


Fig. 2: Throughput of SW, SW-NC and SW-NC-FEC in Error free channel

A. *Performance Analysis*

For the sake of clarity: SW represents Stop-and-Wait Automatic Repeat reQuest; SW-NC represents Stop-and-Wait Automatic Repeat reQuest with Network Coding which is used in [8]; while SW-NC-FEC represents Stop-and-Wait Automatic Repeat reQuest with Network Coding and Forward Error Correction, which is the proposed system in this paper.

Over an error free channel, the maximum achievable throughput for SW, SW-NC and SW-NC-FEC are shown in Figure 2. It is clear that network coding with forward error correction is superior. This can be attributed to the fact that in an error free channel only one transmission is required in order to transmit all the n packets. In the system developed in [8] however, all packets have to be transmitted mindless of whether earlier sent packets have been received successfully. For SW-ARQ, the throughput of the system is around 33 kbps. This is expected because the time required to transmit a block containing n packets in SW-NC-FEC is the same time used by SW-ARQ in transmitting a single packet.

As the frame error rate (FER) increases the delay in transmission of packet increases, hence the difference in performance (i.e. throughput) as shown in Figure 3, 4 and 5. To investigate the severity in drop of performance with increase in FER, a graph of throughput for the three simulations against their channel’s FER (See: Figure 6) is plotted. From the graph, it can be seen that fall in throughput is more obvious in the case

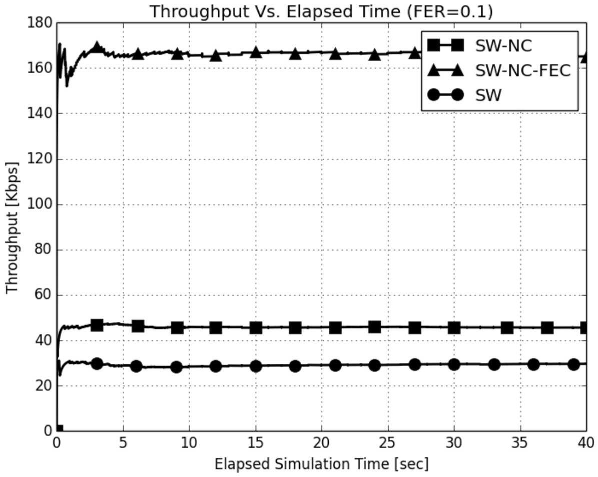


Fig. 3: Throughput of SW, SW-NC and SW-NC when FER=0.1

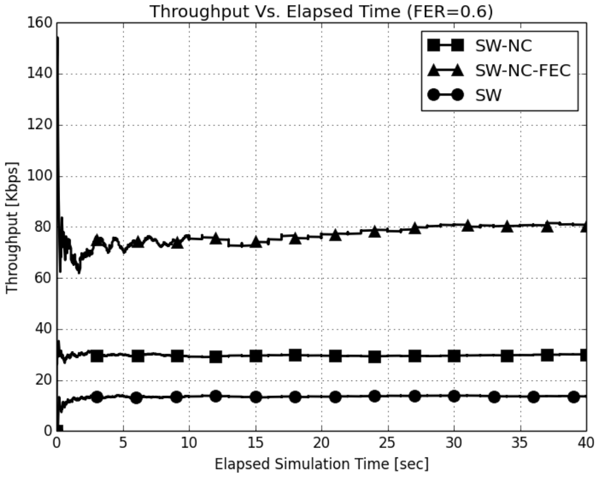


Fig. 4: Throughput of SW, SW-NC and SW-NC-FEC when FER=0.6

of our proposed system. This is due to the fact that number of re-transmissions in SW-NC-FEC approaches the number of re-transmissions of the SW-NC as FER approaches 1. As such, the larger the error rate the more packets SW-NC-FEC needs to transfer and the closer its behavior is to SW-NC in terms of number of packets sent. Thus the cost of packet loss is higher when FEC is added.

In addition, SW-NC-FEC is greatly affected by packet generation rate. This can be seen in Figure 7, where SW-NC-FEC steeply climbs as packet generation rate increases. Conversely, SW-NC-FEC does not perform well with small generation rate. In fact, it performs worse than SW-NC when generation rate is below 0.03 packets/s.

Finally, scalability in terms of increase in number of packets encoded is investigated (see: Figure 8). It is found that SW-NC-FEC's performance increases linearly with increase in number of packets while SW-NC's performance decreases linearly (although slightly) with increase in number of packets.

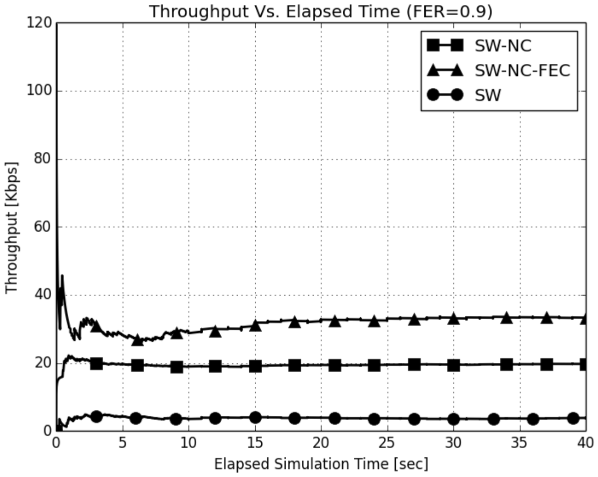


Fig. 5: Throughput of SW, SW-NC and SW-NC-FEC when FER=0.9

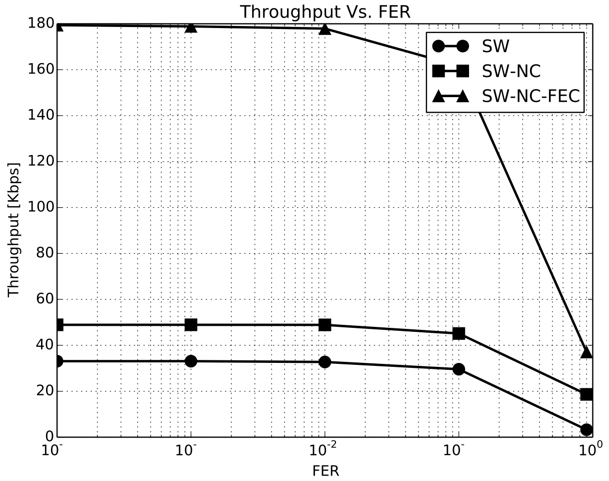


Fig. 6: Throughput of SW, SW-NC and SW-NC-FEC over a wide range of FER

B. Application

Table II summarizes the characteristics of SW, SW-NC and SW-NC-FEC. Base on this comparison, it can be seen that SW-NC shows more stable performance in a noisy environment and preforms better in a very low traffic network. Furthermore, the system is less complex than SW-NC-FEC, as such it shorter code and less processing overhead. Therefore, SW-NC is better for low traffic network systems where low processing power devices are used like wireless sensor networks, while SW-NC-FEC is best for networks that require large amount of data to be transmitted over a low noise channel.

TABLE II: Comparison between SW, SW-NC and SW-NC-FEC

	SW	SW-NC	SW-NC-FEC
Complexity	Simple	More complex	Most complex
Memory overhead	Less	More	Most
Throughput @ FER=0	33.0kbps	48.4kbps (147.0% of SW)	180.0kbps(545.5% of SW)
Throughput @ FER=0.9	6.0kbps	18.0kbps (300% of SW)	36.0kbps(600% of SW)
Decay in throughput with increase in FER	Faster	Fast	Fastest
Throughput as No. of packets encoded increase	-	Falls linearly	Increases linearly

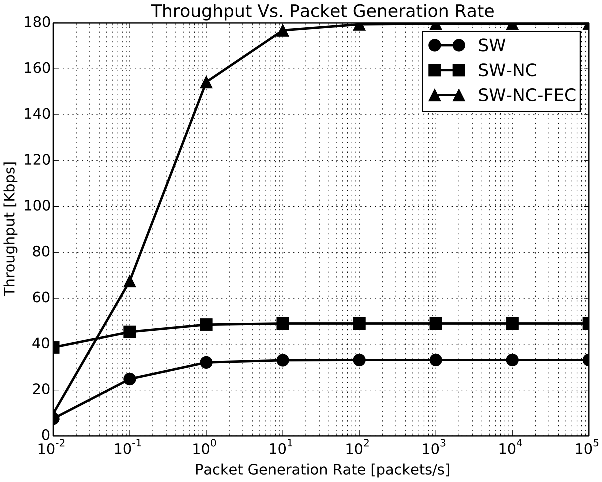


Fig. 7: Throughput of SW, SW-NC and SW-NC-FEC against packet generation rate

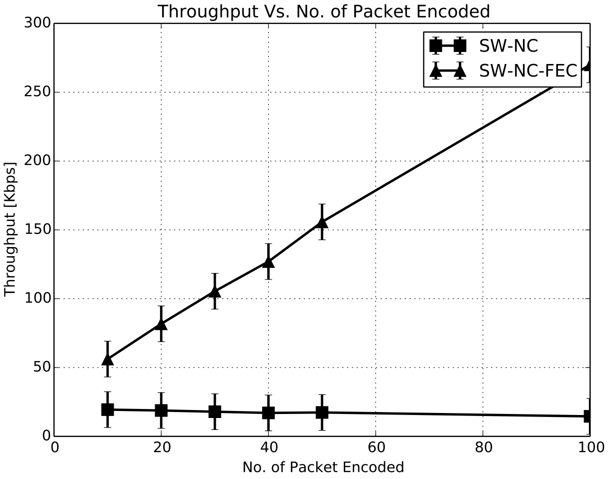


Fig. 8: Throughput of SW-NC and SW-NC-FEC against number of packets encoded

VI. Conclusions and Future Work

In this paper, the throughput of a SW-ARQ using Vandermonde matrix for network coding is analyzed. The study shows that NC with FEC is highly profitable and that adding FEC to the system in [8] improves its throughput. Furthermore, the study show that SW-NC-FEC has higher throughput than SW-NC, but its throughput falls more sharply with increase in frame error rate. Conversely, it is also found that SW-NC-FEC rises more sharply when packet generation rate increase. Finally we found that Forward Error Correction helps SW-NC-FEC to produce more throughput as Vandermonde matrix is enlarged, while SW-NC degrades.

Due to the fact that the connection between the transmitters and the receivers is many-to-one, there is a possibility of performance degradation when the number of transmitters is increased. As their number increases, a bottle neck is may formed at the encoder and this may cause delay in the network. However, the degree of decay in performance of a network-coded network has not been investigated. It is important to ascertain when the throughput starts to degrade and how bad is the degradation, and this forms the basis of our future work.

Acknowledgements

We would like to acknowledge King Fahd University of Petroleum and Minerals (KFUPM) for its continuous support to academic research activities.

References

[1] Y. Qin and L.-L. Yang, "Steady-state throughput analysis of network coding nodes employing stop-and-wait automatic repeat request," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1402–1411, Oct 2012.

[2] M. Zhang, "Major automatic repeat request protocols generalization and future develop direction," in *6th International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII)*, vol. 2. IEEE, 2013, pp. 5–8.

[3] R. Ahlswede, N. Cai, S.-Y. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[4] J.-Y. Lee, W.-J. Kim, J.-Y. Baek, and Y.-J. Suh, "A wireless network coding scheme with forward error correction code in wireless mesh networks," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE.* IEEE, 2009, pp. 1–6.

[5] A. A. Bruen and M. A. Forcinito, *Cryptography, information theory, and error-correction: a handbook for the 21st century.* John Wiley & Sons, 2011, vol. 68.

[6] C. Fragouli, J.-Y. Le Boudec, and J. Widmer, "Network coding: An instant primer," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 1, pp. 63–68, Jan. 2006. [Online]. Available: http://doi.acm.org/10.1145/1111322.1111337

[7] S. Luyi, F. Jinyi, and Y. Xiaohua, "Forward error correction," in *Fourth International Conference on Computational and Information Sciences (ICCIS).* IEEE, 2012, pp. 37–40.

[8] A. Alsebae, M. Leeson, and R. Green, "The throughput benefits of network coding for sw-arq communication," in *27th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, March 2013, pp. 854–859.

[9] S. De Vuyst, S. Wittevrongel, and H. Bruneel, "Delay analysis of the stop-and-wait arq protocol over a correlated error channel," in *Proc. of the second international working conference on performance modelling and evaluation of heterogeneous networks HET-NETs' 04*, 2004, p. 21.

[10] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell system technical journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[11] Q.-T. Quoc-Tuan Vien, L.-N. Tran, and H. X. Nguyen, "Network coding-based arq retransmission strategies for two-way wireless relay networks," in *Software, Telecommunications and Computer Networks (SoftCOM), 2010 International Conference on*. IEEE, 2010, pp. 180–184.

[12] A. Antonopoulos and C. Verikoukis, "Network coding based cooperative arq scheme," *arXiv preprint arXiv:1201.4650*, 2012.

[13] B. Li and D. Niu, "Random network coding in peer-to-peer networks: from theory to practice," *Proceedings of the IEEE*, vol. 99, no. 3, pp. 513–523, 2011.

[14] Y. Liu, Z. Feng, and P. Zhang, "A novel arq scheme based on network coding theory in cognitive radio networks," in *IEEE International Conference on Wireless Information Technology and Systems (ICWITS)*. IEEE, 2010, pp. 1–4.

[15] Y. Qin and L.-L. Yang, "Throughput analysis of stop-and-wait automatic repeat request scheme for network coding nodes," in *Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st*. IEEE, 2010, pp. 1–5.

[16] X. Chen and D. Leith, "Frames in outdoor 802.11 wlans provide a hybrid binary-symmetric/packet-erasure channel," *arXiv preprint arXiv:1209.4504*, 2012.

[17] N. Ilievska and D. Gligoroski, "Error-detecting code using linear quasigroups," in *ICT Innovations 2014*, ser. Advances in Intelligent Systems and Computing, A. M. Bogdanova and D. Gjorgjevikj, Eds. Springer International Publishing, 2015, vol. 311, pp. 309–318. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09879-1_31

[18] A. Al-Shaikhi and J. Ilow, "Vandermonde matrix packet-level fec for joint recovery from errors and packet loss," in *Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on*, Sept 2008, pp. 1–6.

[19] G. S. Fishman, *Principles of discrete event simulation.[Book review]*. John Wiley and Sons, New York, NY, 1978.

[20] D. P. Kroese, T. Taimre, and Z. I. Botev, *Discrete Event Simulation*. Wiley Online Library, 2011.

[21] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science and engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[22] S. Vaingast, "The environment," in *Beginning Python Visualization*. Apress, 2014, pp. 31–53. [Online]. Available: http://dx.doi.org/10.1007/978-1-4842-0052-0_2