IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 6 Issue 10

October 2015

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

SAI

www.ijacsa.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 10 October 2015
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Framework for Assessing Privacy of Internet Applications

James PH Coleman
Department of Computing
Edge Hill University,
Ormskirk, Lancashire, United Kingdom

***Abstract*—This paper presents a new framework for assessing and documenting the privacy risks associated with developing and managing internet applications. The Framework for Assessing Privacy of Internet Applications (FAPIA) provides a tool to aid the analysis of privacy risks and a structured means of analyzing the risks and documenting a control systems to ensure compliance with data protection and privacy legislation in a range of different countries.**

Keywords—privacy; privacy compliance; risk; data protection; privacy impact assessments; internet applications

I. Introduction

The “Internet of Things” [1], “Ubiquitous Computing” [2] and “Private area networks” [3] are similar concepts and relate to computing and network devices that are being created and installed to improve life and lifestyle. There are CCTV cameras in railway stations and along busy shopping streets, there are RFID tags to protect expensive clothing in high street shops as well as wearable health technology, eg devices to monitor heart beat with the ability to “send” this data to other devices and people.

Care and attention is being paid by developers to building user friendly interfaces, and secure encryption of data over networks, what is lacking is that far more subtle of concept – that of privacy. Internet users are concerned about the privacy of their personal information when it is online for it can lead to invasions of privacy [4], [5]. A study in 2004 [6] shows that users are particularly concerned about 3 aspects:

- the collection of personal information,
- the control of personal information and
- the awareness of users about actual privacy practices being used by system developers.

Given the wide range of applications that are involved, including: home and private networks, smart metering, monitoring for health and social care needs, observation by CCTV and other tools of public spaces, airports and train stations as well as the increasing use of the public Internet for phone calls and emails etc. with tools able to “*provide tailored services*” based on the content of emails etc. It is increasingly important that users and developers are aware of the privacy implications of their existing and any new systems or services they offer and experience tells us that a powerful tool for understanding this is the privacy Impact Assessment.

As the International Association For Impact Assessment (IAIA) [7] state - an “*... Impact assessment ... is the process of identifying the future consequences of a current or proposed action*”.

In this article, a framework is proposed that will help developers with the understanding of the potential privacy implications of their system. It aims to aide developers and system managers to:

- ensure compliance with relevant privacy and data protection laws,
- manage risks to the perception of the system’s owner and consumer confidence in the system (specifically related to privacy and data protection), and
- develop confidence in these new technologies

As privacy and data protection is applicable to a wide range of different applications and systems, it will provide a set of tools that developers and managers can use.

A Privacy Impact Assessment (PIA) is a document which “…seeks to… [explicitly].., in as much detail as is necessary…, [specify]…the essential components of any personal information system or any system that contains significant amounts of personal information…[under the headings]… description; data collection; disclosure and use of data; privacy standards and security measures…” [8] This will help developers to fully understand the privacy and data protection implications of the system they are developing, and to then be able to provide the necessary and relevant information that users, customers or the public want and need about the systems that will be used or are proposed. The PIA is a very generic concept that needs to be adapted to differing situations and scenarios, not just those relating to networked computerised environments. This article will focus on its applicability to networked systems, (where the Internet is involved) and will provide a framework for developers to frame their PIA analysis and develop better networked systems.

A. Definitions of a PIA

There are numerous definitions of a PIA and while the definitions are different, they agree that the PIA is an “assessment of any actual or potential effects that the activity or proposal may have on individual privacy and the ways in which any adverse effects may be mitigated” [9]

"PIA is an analysis of how information in identifiable form is collected, stored, protected, shared and managed. [to] ensure that system owners and developers have consciously incorporated privacy protection throughout the entire life cycle of a system" [10]

''a process whereby the potential impacts and implications of proposals that involve potential privacy-invasiveness are surfaced and examined'' [11]

B. Framework for assessing privacy

The purpose of this Framework is to assess the privacy risks involved in a computerised system, and having determined the risks to privacy, to then enable the system developers/managers to ensure those risks are properly managed. The precise systems or controls a developer uses to manage those risks will depend upon the legal framework in which they operate however this Framework will enable them to understand what is happening (from a privacy perspective) in their system.

By working through the Framework the developer/manager will be able to develop controls (either technological or procedural) to manage privacy within the system.

The Framework for Assessing Privacy for Internet applications (FAPIA) is a tool that is designed to simplify the PIA process for internet application developers. It will enable them to assess the privacy risks, assess the likelihood and provide a structured way of addressing these risks and providing a documented outcome, particularly if a legislative framework exists, in which the system will operate. In the words of [8] "*... a basic function of a ... privacy impact assessment is to ask probing, detailed questions of the proponents, builders and designers in order to promote comprehension. The role is in effect that of a devil's advocate*."

FAPIA has been developed in the UK and therefore has been heavily influenced by the needs of the European Convention on Human Rights [12], in particular Article 8 which states:

Article 8 – Right to respect for private and family life

1) Everyone has the right to respect for his private and family life, his home and his correspondence. [12 Art 8]

This is a limited right as it goes on to say in section 2 - that *"…there shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society…"* and goes on to define certain conditions when this interference can occur.

Related to this is the European Union (EU) Directive on Data Protection [13] which provides a set of 8 principles on the collection and processing of personal data. While this may seem to be European-centric, it must be remembered that many nations have privacy laws or statutory requirements surrounding the processing of data and privacy including the so-called "safe harbour" schemes of the US Department of Commerce [9] which provides 7 principles for data processing and storage that must be used in certain circumstances.

While traditionally privacy relates to proper persons – that is, actual identifiable people. As the Internet grows however, and as the number of institutions uses digital systems to manage their data and environment, it is important that privacy be considered in a wider context. Under traditional terminology, privacy relates to a person, in the modern world we need to consider privacy in relation to institutions and organisations. Businesses use digital and internet systems to store their data and they expect this information (perhaps about business costs, incomes etc) to be properly managed. Managed in terms of data loss but also in the form of being kept confidential. Whole industries are reliant on being able to control access to business data (which may not fall under the protection offered by privacy legislation). Therefore while legal privacy protection may not exist, privacy as a concept applies equally to business data as it does to personal data given contractual obligations that form part of many business contracts.

As the author is EU based, then FAPIA will ensure that ECHR/DFD [12, 13] can be implemented, it does mean that FAPIA is equally usable and valid in other legislative areas and the Author believes is valid for all existing jurisdictions. This article will explain how FAPIA is structured, and how it can be applied to different domains.

II. Framework for Assessing Privacy for Internet Applications

The first question that needs to be asked by a developer is whether there are any privacy implications. This relates to whether issues of privacy as a result of legal or contractual obligations exist. Figure 1 for a decision tree which is the first step in the FAPIA process.

An assessment completed using FAPIA will consider each of the elements of the Internet application (called the Key Elements in FAPIA) and will analyse these in relationship with the privacy targets (called Privacy Targets in FAPIA) and will identify what FAPIA calls the Key Outcomes.

FAPIA is a system that aims to help developers understand their own (ideally proposed) system, which will then allow the developers to build a system that has minimal privacy impacts, or at least where the impacts can be managed properly and easily. Often what happens is that Developers find that they have built a system that is difficult to control, or where expensive "*procedural*" systems or add-ons are needed to maintain the level and type of privacy and data protection that is needed and expected from the customers.

Having applied the FAPIA, Developers should then have an understanding of the implications of what they are proposing, and are able to build the relevant controls into the system. It is unfortunate that many developers do not consider the privacy impacts until the end of the design phase (and sometimes the end of the development phase) of a system when modifications can be both expensive and difficult.

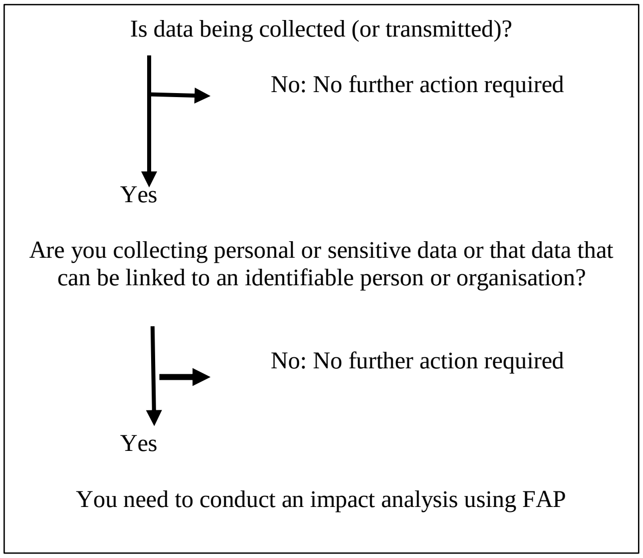

Is data being collected (or transmitted)?
No: No further action required
Yes
Are you collecting personal or sensitive data or that data that can be linked to an identifiable person or organisation?
No: No further action required
Yes
You need to conduct an impact analysis using FAP

Fig. 1. Decision Tree for deciding if an assessment is required

A. Key Outcomes

The *Key Outcomes* of FAPIA are the following pieces of information about how the IA. The

- **Business need** being addressed by the proposed system?
- The **data being collected**, and the reason why that data is being collected?
- Description of the **Mechanisms** used in the IA which processes or stores (even temporarily) data?
- What **privacy impacts** may arise as a result of?
  - Algorithms used to process the data
  - Data storage and is that storage necessary for the operation of the system?
  - What is the justification for negative privacy impacts?
- **Procedures** and protocols used to manage access to data
  - What methods are proposed whereby those negative privacy impacts will be ameliorated?

In order to identify the Key Outcomes of an IA, it is first necessary to identify main actors/agents that are involved. These are called the Key Elements and it is the analysis of what is generally pieces of software and/or hardware that will result in improved understanding of the privacy impact of an IA.

B. Key Elements

The Key Elements identifies the components that Developers and Maintainers need to identify in their system. By doing so they are then able to develop and maintain the systems and controls that are needed to maintain privacy for not all risks can be addressed by technology alone.

The Key Elements are:

Data Subjects – The Data Subject is the person (or group of people) who is the subject of the collection of data. Note that this may not be the same as the intended data subject.

The Data Subject is the individual whose particular data is provided to the system and the person can be identified or distinguished from others in the system. (eg a person has brought a product that uses Radio-frequency identification (RFID) tags to identify products in a store, and the same RFID readers are used in a number of shops, so by linking the original sales transaction to the RFID ID then the person can be identified).

In identifying the data subjects, it is important to consider those who are not the intended subjects. So for example, in a CCTV system for a railway station, all people who move through the station may be data subjects if the images are recorded (for 14 days) in a crime-detection programme. Often the data subject is not the person about which data is being generated but is the person who receives the data. So for example, in an email system the data subjects are both the recipients and senders of the emails.

Data Elements – This is the data that is being collected and which may be processed in response to instructions and/or is recorded with the intention that it should be processed at a later date. The data is collected (or being sent) by the Network Device. In determining the data elements involved in the IA, it is important to remember that often other data is being collected as well as that which is intended.

WHAT IS PERSONAL DATA?

The data collected, processed and/or stored – It is important to realise that privacy is primarily concerned about data that is related to a person or a group of people or which can be linked to a real person or organisation. This means that if the system is collecting data on activities or events for which the data is not linkable to an identifiable person/organisation, then privacy considerations are generally not applicable. So for example, a RFID system does not need to consider privacy considerations when the tags are in the warehouse, and there is no link to a person. Once the link to a person can be made because of a sale, then privacy considerations now apply.

Network Sensor – A network sensors is a device to monitor conditions (eg CCTV, temperature, sound, pressure, etc). and to pass their data through the network to a main location – an Application. As networks may be bi-directional, also enabling control of sensor activity This means that the network sensor may be a movement sensor, biometric sensor or a keyboard or tablet computer.

Network Node – The network node includes devices that pass on data received from a sensor or pass on data over a network (which may include the public Internet but is quite often a private network or a combination). There may be multiple network nodes involved in data collection (or transmission).

Application – An application is a process or group of processes that involve the processing of the data collected about a data subject which includes potentially the linking of this data with other data sources. This can include the organisation, retrieval, disclosure of data or the linking of collected data with other data sets.

Back-end – The back-end is the area used for the storage of data about or involving a data subject. This includes the ability to store what is initially un-linked data but through later analysis linkages are created.

Data Controller – the Data Controller is a person or an organisation or a group who process or use the data about a data subject. This includes the situation where an organisation is processing (including storing) data on behalf of another organisation in the form of an out-sourcing arrangement.

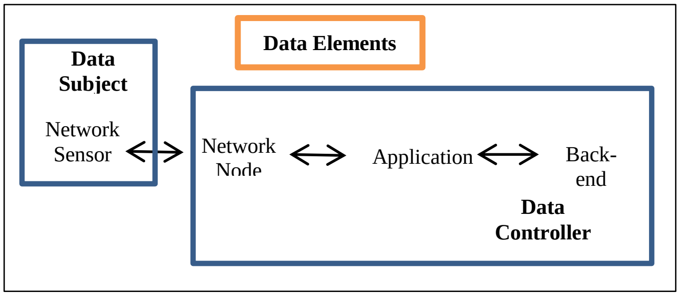


Fig. 2. FAPIA Key Elements

C. Privacy Targets

In order to develop a system that is compliant with privacy legislation (as implemented in different regions of the world), it is necessary that the developers fully understand their own system, who is involved in the system, and how it works. Once the Key Elements have been identified, FAPIA then tests these elements against privacy and data protection targets, called the *Privacy Targets*. These are characteristics that any system that manipulates personal or sensitive information needs to address. It should be borne in mind that different countries will have precise legal definitions for some of these characteristics but this Framework is structured in such a way that the information will be available for most legislative systems and that the developers will have the information necessary to provide reassurance for commercial companies and organisations about the privacy services that will be provided.

The Privacy targets identified as part of this Framework are based on the pargets identified as part of the EU Privacy Directive. [13].

These targets are grouped according to the 3 issues identified by [6] as being the main issues that network users are concerned about when considering privacy. The targets represent the concerns of users or customers and as such should represent a major driver for businesses who use data for whatever reason.

These targets are:

1) Collection of Private Information

Safeguarding quality of private data

Data avoidance and minimisation, purpose specification and limitation, quality of data and transparency are the key targets that need to be ensured.

Legitimacy of processing data

Legitimacy of processing personal and private data must be ensured either by basing data processing on consent, contract, legal obligation, etc.

2) Control of personal information

Compliance with the data subject's right to be informed

It must be ensured that the data subject is informed about the collection of his data in a timely manner.

Compliance with the data subject's right to object

Where there is a legal right to object (or *right-to-be forgotten*) to having data processed, then this is ensured. Transparency of automated decisions vis-à-vis individuals must be ensured especially. Even if there is no *right to object*, the Developers need to provide a mechanism for handling Data Subject's expression of views and how this is to be handled by the system. All existing privacy frameworks include some mechanism for appeals

Safeguarding confidentiality and security of processing

Preventing unauthorised access, logging of data processing, network and transport security and preventing accidental loss of data are the key targets that need to be ensured.

Allowing the data subject right of access to data, and to correct and erase data

It must be ensured that the data subject's wish to access, correct, erase and block his data is fulfilled in a timely manner. All existing privacy frameworks include some mechanism for appeals

3) Awareness of users about actual privacy practices

Compliance with notification requirements

Notification about data processing, prior compliance checking and documentation are the key targets that need to be ensured.

Compliance with data retention requirements

Retention of data should be for the minimum period of time consistent with the purpose of the retention or other legal requirements.

III. Outcomes and Future Work

As FAPIA is a process which generates a series of outcomes (see section II. A. above) this results normally in the production of a report. If the Framework was applied prior to the development of the system itself, then this report provides input into the design process for the Internet application.

Where other tools need to be addressed, then FAPIA provides a comprehensive data set that can be used. FAPIA however produces a set of Key Outcomes:

Overview specification of the Operation of the system – the business needs being addressed.

Specification of actual data collected – a comprehensive set of data elements being collected explaining how the data is collected and the purpose being addressed by the data.

Specification of actual data stored – a comprehensive set of data elements being stored explaining how the data is stored and the purpose being addressed by the data storage. This also needs to include how the data is being protected given the level of the risk involved. This also needs to include temporary storage of data (eg hours or days).

Algorithms used to process the data – this is important to know how the system is processing the data. As a consequence of data processing, some data which was collected anonymously may become identifiable, and hence become a privacy risk that needs to be managed.

Privacy Risks – what are the privacy-risks associated with the collection, processing and storage of the data. Where there are privacy risks then the justification for having negative privacy risks need to be provided, depending on the level of the risk-score involved.

Privacy Systems - Amongst the numerous different types of privacy-risks, privacy systems/processes will be used to manage the privacy requirements of all systems. There are a number of *common procedures and protocols* which all systems need to address:

- Procedures and protocols used to manage access to data and other Key Elements
- Procedures and protocols for the collection of data
- Procedures and protocols for informing the data subjects
- Procedures and protocols for managing data transfers
- Procedures and protocols for managing data loss
- Procedures and protocols for objections, appeals
- Procedures and protocols for providing data subject to access, correct and erase data (where needed or necessary)

Figure 3 illustrates a form of FAPIA Report for a sample CCTV system which consists of a video camera with images being stored on a PC. The FAPIA Report

Having completed FAPIA and produced these outputs, the system developers and managers have a very comprehensive understanding of how their system needs to be developed. The system managements tasks that need to be undertaken, and often can be built into the project are clearly identifiable within the FAPIA key Outputs.

FAPIA Report – CCTV Monitoring System

Key Elements:
Data Controller: Sample Organization Owner
Data Subjects:

- Customers entering Sample Organization's shops.
- Members of staff of Sample Organization who work in areas covered by CCTV system.
- Members of public who walk in front of sample Organization's shop.

Network Sensor:	CCTV camera in shop
Network Node:	PC where *CCTVSoftware* is located
Application:	*CCTVSoftware* application
Back-end:	*CCTVSoftware* and DVD storage

Key Outputs:

Overview specification of the Operation of the system The CCTVSoftware system consists of a Camera and related software which records all images the camera captures and stores these images for a 7 day rotational system. The camera captures images of activity within the shop-area including through the window and on the street front.

Specification of actual data collected
Images of members of public, customers and staff in the shop or outside front shop window.

Specification of actual data stored
Images of members of public, customers and staff in the shop or outside front window stored for 7 days on rotational cycle unless retained for a longer period for crime detection/prevention purposes

Algorithms used to process the data
The only algorithms used are in the capture and storage of data. Data used for crime detection/prevention purposes may be enhanced to provide better images for forensic purposes.

Privacy Risks

- Use of material for non-approved purposes
- Loss of data – possible because of:
  - Technical problems (eg data storage shortage)
  - Procedural problems (eg staff deleting files)

Privacy Systems - Procedures and protocols used to:
Manage access to data and other Key Elements

- Procedure for examining images in the event of suspected crimes

Procedures and protocols for informing the data subjects

- CCTV Notice – Notice informing the public of the use of CCTV for crime detection/prevention uses.

Procedures and protocols for the collection of data

- Protection of stored data in locked room

Procedures and protocols for managing data transfers

- Procedure for storage and copying of images in the event of suspected crimes

Fig. 3. Sample FAPIA Report

A. Other Outcomes

FAPIA can also be used to provide the information needed for a range of other reports, especially systems such as Privacy Impact Assessments as is required by a number of organisations as a result of government regulations. While the precise structure and content of the PIA will vary, often markedly, the comprehensiveness of the FAPIA system will mean that in most cases the information needed has been collected by the FAPIA process and is readily available.

The main outputs of the FAPIA process is an assessment of the impact on privacy of an internet application. This set of impacts could be used to inform part of the risk register for the system. The Risk Register (RR) is a risk management tool commonly used in risk management and compliance. It acts as a central repository for all risks identified by the organisation. Although risk registers are used extensively they are often criticised because they can lead to ritualistic decision-making [14], illusion of control [15], and the fallacy of misplaced concreteness [16]. Used correctly however they can be a useful tool in managing the risks associated with a project. This suggests that a similar tool would prove equally useful in managing the privacy concerns of applications

IV. Conclusion

The Framework for the Assessment of Privacy of Internet Applications (FAPIA) is a tool that if used, enables Internet system developers to build systems that respect the privacy of the people who are involved in the system while still achieving the business needs for which the system was developed and should do so without having to bolt-on systems after completion (or near completion). FAPIA focusses on Internet systems, and as such provides specific internet-linked prompts that developers can use.

References

[1] The Internet of Things Council, http://www.theinternetofthings.eu/, Accessed: 9/5/2015.

[2] Weiser, M., Brown, J.S.: The coming age of calm technology. (1996) http://www.ubiq.com/hypertext/weiser/acmfuture2endnote.htm. Accessed: 29/5/2015

[3] "Personal Area Networks", http://www.techopedia.com/definition/5079/personal-area-network-pan, Accessed: 9/5/2015.

[4] Laufer, R. S., & Wolfe, M, "Privacy as a concept and a social issue: A multidimensional developmental theory". Journal of Social Issues, 33, 1977, pp22–42

[5] Culnan, M. J., "Protecting Privacy Online: Is Self-Regulation Working?," Journal of Public Policy and Marketing (19:1), 2000, pp. 20-26

[6] Malhotra N., Kim S., Agarwal J. "Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model", Information Systems Research, Vol 15, Issue 4, 2004, pp336 – 355.7 6-0a International Association For Impact Assessment

[7] IAIA. International Association For Impact Assessment, http://www.iaia.org/. Accessed: 9/5/2015

[8] 8# 6a Flaherty, D., "Privacy impact assessments: an essential tool for data protection" [2000] PrivLawPRpr 45; (2000) 7(5) Privacy Law and Policy Reporter 85; http://www.austlii.edu.au/au/journals/PLPR/2000/45.html Accessed: 9/5/2015

[9] 9# 6a1 Stewart, B., "Privacy impact assessments" [1996] PrivLawPRpr 39; (1996) 3(4) Privacy Law & Policy Reporter 61. http://www.austlii.edu.au/au/journals/PLPR/1996/39.html. Accessed: 9/5/2015

[10] 10# 6a2 Clarke R., "Privacy impact assessment: Its origins and development" computer law & security review 25, 2009, pp123–135

[11] 11# 6a3 Clarke R. "Privacy impact assessments", 1998, Xamax Consultancy Pty Ltd. Available at: http://www.xamax.com.au/DV/PIA.html. Accessed: 9/5/2015

[12] "European Convention on Human Rights", Council of Europe, http://www.echr.coe.int/Documents/Convention_ENG.pdf, Accessed: 9/5/2015

[13] "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of Individuals with regard to the processing of personal data and on the free movement of such data," Official Journal of the European Communities, 1995, L 281/31, http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf. Accessed: 9/5/2015

[14] Drummond, H. "MIS and illusions of control: an analysis of the risks of risk management. Journal of Information Technology, 2011, 26, 259–267. doi:10.1057/jit.2011.9

[15] Lyytinen, K. "MIS: the urge to control and the control of illusions – towards a dialectic". Journal of Information Technology, 2011, 26, 268-270 doi:10.1057/jit.2011.12

[16] Budzier, A. "The risk of risk registers – managing risk is managing discourse not tools". Journal of Information Technology, 2011, 26, 274-276 doi:10.1057/jit.2011.13

Appendices

The Data Protection Directive [13] of the EU provides the following set of principles for the processing of data and related principles of privacy. These principles underpin this Framework.

Data protection principles

1) Personal data shall be processed fairly and lawfully.

2) Personal data shall be obtained only for specified and lawful purposes, and shall not be further processed in any manner incompatible with that purpose or those purposes.

3) Personal data shall be adequate, relevant and not excessive in relation to the purposes for which they are processed.

4) Personal data shall be accurate and, where necessary, kept up to date.

5) Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes.

6) Personal data shall be processed in accordance with the rights of data subjects.

7) Appropriate technical and organisational measures shall be taken against unauthorised or unlawful processing of personal data and against accidental loss or destruction of, or damage to, personal data.

8) Personal data shall not be transferred to a country or territory outside the EU unless that country or territory ensures an adequate level of protection

Towards Network-Aware Composition of Big Data Services in the Cloud

Umar SHEHU
Department of Computer Science and Technology University of Bedfordshire Luton, UK

Ghazanfar SAFDAR
Department of Computer Science and Technology University of Bedfordshire Luton, UK

Gregory EPIPHANIOU
Department of Computer Science and Technology University of Bedfordshire Luton, UK

***Abstract*—Several Big data services have been developed on the cloud to meet increasingly complex needs of users. Most times a single Big data service may not be capable in satisfying user requests. As a result, it has become necessary to aggregate services from different Big data providers together in order to execute the user's request. This in turn has posed a great challenge; how to optimally compose services from a given set of Big data providers without affecting if not optimizing Quality of Service (QoS). With the advent of cloud-based Big data applications composed of services spread across different network environments, QoS of the network has become important in determining the true performance of composite services. However current studies fail to consider the impact of QoS of network on composite service selection. Therefore a novel network-aware genetic algorithm is proposed to perform composition of Big data services in the cloud. The algorithm adopts an extended QoS model which separates QoS of network from service QoS. It also uses a novel network coordinate system in finding composite services that have low network latency without compromising service QoS. Results of evaluation indicate that the proposed approach finds low latency and QoS-optimal compositions when compared with current approaches.**

Keywords—Big data; Service composition; QoS; Genetic Algorithm; Network latency; Cloud

I. Introduction

Service Oriented Computing (SOC) is a framework that allows for internet applications to be built by coupling web services together. In SOC, each web service represents a different functional aspect of a Service-oriented application (SOA) [32].

Web services are network-accessible objects that allow Big data vendors to build service-oriented Big data applications (SOA) which share business logic and application services with other vendors in order to meet growing consumer needs. A Big data service (BDS), also known as Big-data-as-a-service (BDaaS), is a data intensive web service that works on large scale unstructured or semi-structured datasets. They typically perform tasks such as data storage, processing, cleaning, extraction, modelling and virtualization on large datasets. BDaaS consist mainly of three layers namely; the infrastructure layer, platform layer and application layer. The infrastructure layer provisions the physical resources required to process large datasets. The platform layer houses the operating systems and virtual machines that run BDaaS applications. The application layer represents the models and software used to process Big data.

Every BDS is characterized by the ability to provide some task as identified by its functional and non-functional attributes [27]. The functional attributes define what the service is capable of doing e.g. Microsoft Azure BDS [30] provides cloud-based machine learning framework for analyzing large scale datasets. Non-functional attributes on the other hand determine how well a service can perform a given task e.g. how long it will take Microsoft Azure BDS to respond to a user request. Non-functional attributes are commonly referred to as QoS (Quality of Service). Examples of service QoS attributes include response time, cost, reputation, etc. They are often used as criteria in selecting services suitable for a user request especially in situation where there is more than one service with similar functionality. For instance, a Microsoft Azure BDS having response time equal to 10 seconds will be more suitable to a user requiring prompt service response than a similar service such as Amazon AWS BDS with a response time of 30 seconds. Thus service QoS is used to differentiate services that are similar in terms of their functionalities. In SOC, functionally similar services are usually categorized in the same service group and referred to as candidate services.

A. QoS-Aware Web Service Composition

Recently, the ability of services to aggregate their functionalities has gained much attention. This is as a result of increased complexity of user requests. Simple user requests may require only a single BDS to be completed. However, as user requests take more complex forms that are beyond the capabilities of a single BDS, aggregating service abilities is necessary to carry out the request. This process is known as service composition. It combines services in order to build a composite service [8, 9] that is viewed by the user as a single service. Within a composite service each constituent BDS takes care of a specific functional aspect of the user’s request. For instance suppose a user issues a compound request like "Analyse e-books dataset" consisting of several sub requests at the task level such as Twitter feed analysis, Natural language processing and IOT device log analysis (as seen in Fig. 1). A single BDS is ill sufficient to satisfy the compound request, therefore services from different Big data vendors for each sub request will need to be discovered and aggregated together according to their QoS to complete the user request. QoS-aware service composition process is similar in principle to the behaviour of workflow management system [28] in which a workflow dictates how data should be processes. A workflow processes data by using different patterns that transform data to an end result. Service composition uses similar patterns found in workflow systems. The patterns allow composition process to channel data flow from one BDS to another. Some major service composition patterns include sequence, parallel, exclusive choice and loop. Service composition process begins by breaking down a complex request into smaller functions or sub-requests organized according to one of several patterns. Depending on the pattern involved, service QoS are then orchestrated to determine QoS for a composite service. Usually there are several candidate services that exist within a service group that can execute a given functional aspect of the request.

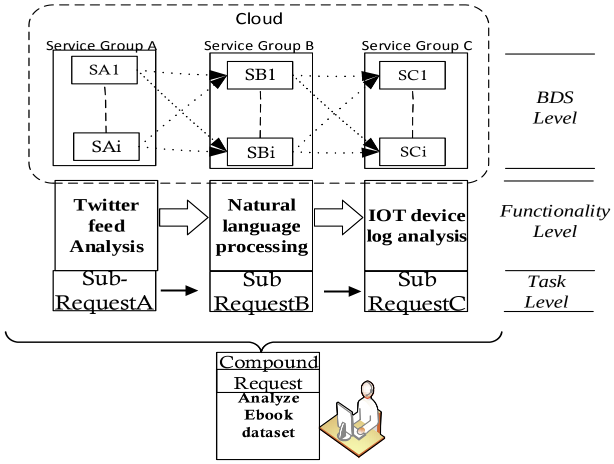


Fig. 1. Typical BDS composition scenario

Therefore choosing a service from each service group that maximizes the QoS of composite service while satisfying the user's constraints has become a research problem. The problem is also known as an NP-Hard optimization problem [18]. The problem has been solved using several techniques such as Linear Integer Programming [13] and Dynamic Programming [11]. Although techniques based on genetic algorithms [18] are usually used in finding near optimal compositions in polynomial time.

B. *Service Composition in the Cloud*

More and more BDS are increasingly being deployed on the cloud with the purpose of allowing Internet users from around the globe to access their functionalities for analysing large datasets. For instance organizations such as Amazon and Microsoft offer public cloud services using Amazon Web Services (AWS) [31] and Windows Azure [30] cloud platforms respectively. These services are deployed on cloud data centres via virtual machines (VM) where consumers can access them from literally any part of the world. VMs enable the processing resources such as CPU, storage and network resources needed to properly run cloud-based BDS. Traditionally, service providers deploy their VMs across several cloud data centres located in different geographical areas to host their BDS. Hence, each user will experience different network performances depending on the geographical location of the hosted service. Thus, when a user tries to invoke a composite service with candidate services spanning different cloud locations, the composition may not be able to deliver on the network performance needs of the user even if it has optimal service QoS. This is because the optimal service QoS only represents application level performance of a composite service but it does not account for its network performance. The impact of the network is usually quantified using a metric such as network latency [22]. The effect of network latency on application performance is noticeable in cloud environments where there is high degree of service distribution. Despite this, current studies do not separate QoS of network from service QoS. Hence, they may produce compositions that have sub-optimal performance when invoked by the user. An example is illustrated in Fig. 2. The example shows several BDS deployed on different clouds. Assuming each cloud consists of two or more BDS and is separated from other clouds by different round trip times (RTT). Also assuming a user request consists of a sequence pattern of the three tasks (t_1, t_2, and t_3) in Fig. 1, with each task having a set of candidate services and their respective QoS scores for cost (*P*), response time (*RT*) and execution time (*ET*). Current approaches will ordinarily pick the QoS optimal composite service (highlighted using bold boxes in Fig. 3) consisting of services (|S_{A1}-S_{B1}-S_{C2}|) with respect to cost, execution time and response time.

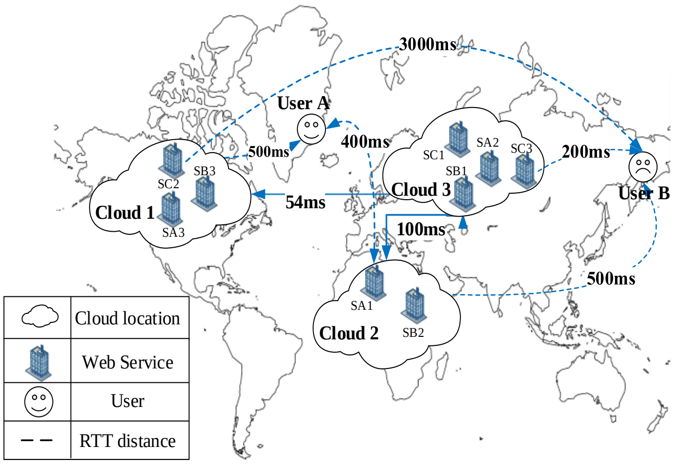


Fig. 2. BDS deployment locations

In doing so, users may experience different levels of performance for this optimal solution depending on the RTT between clouds of participating services.

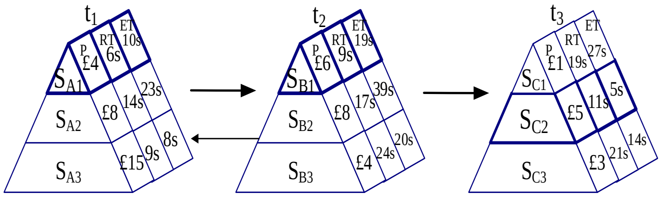


Fig. 3. Sequence workflow pattern with services and their QoS scores

BDS having shorter RTT will incur lower latency than those further away from each other. Therefore user A may experience low network latency for composite service |S_{A1}-S_{B1}-S_{C2}| (i.e. end-to-end network latency for |S_{A1}-S_{B1}-S_{C2}| is 400ms + 100ms + 54ms + 500ms = 1054ms), while user B experiences high network latency because of larger RTT (i.e.

500ms + 100ms + 54ms + 3000ms = 3654ms). Perhaps similar composite services like |S_{A2}-S_{B1}-S_{C2}| (3087ms) or |S_{A2}-S_{B1}-S_{C3}| (311ms) may be better suited for user B since they have lower network latency (as seen in Fig. 3.). This work differs from current approaches in that it separates QoS of network from service QoS. Integrating network latency property into the QoS model will allow us to find composition who's QoS in not only optimal at the application level, but also has near-optimal QoS of network from the user's perspective.

In this paper, a network aware approach to service composition which optimizes network latency and service QoS objectives such as cost, response time and execution time is proposed. It consists of a novel network model which first estimates network latency between BDS in the cloud. Estimation is necessary as traditional latency measurement methods which involve distribution of RTT pings to directly measure RTT between services are generally slow and computationally expensive [4, 6]. Information from the network model is fed to a novel network-aware composition technique based on genetic algorithm in order to find solutions that have optimal service QoS without compromising QoS of the network.

The paper is organized as follows: In Section II an analysis of recent research efforts is presented after which the proposed approach is described in Section III. Section IV presents a discussion of evaluation results. Finally, Section V concludes the paper.

II. Related Work

A. QoS-Based Service Composition

QoS-aware service composition problem has been modelled as an NP-Hard problem [24]. Several classes of approaches have been developed to address the problem. Earlier studies devised local optimization methods to finding optimal composite services. These methods employ search techniques to find services local to each subtask, then combines them into a composite service that will complete the user's request. Techniques developed include dynamic programming [11], learning depth first search [10] and simple additive weighing methods [12]. Another class of approaches widely used are linear integer programming techniques [13, 14]. These techniques use integer variables to search for optimal solutions without having to construct all possible combinations. Meta-heuristic (MH) approaches have been developed to tackle the NP-hard problem. These approaches are based on evolutionary concepts in nature. Some major MH approaches include Genetic [1, 2], particle swarm optimization methods [15] and artificial immune algorithms [29]. All these classes of approaches use similar QoS model which does not take QoS of network into consideration. In comparison, the network-aware genetic algorithm incorporates QoS of the network in the QoS model as it tackles the NP-Hard problem. Genetic algorithm has been chosen because it shows great promise in solving constrained multi-objective optimization. It is also capable of producing a set of solutions in which no solution is dominant to the others, thereby giving the user a wide range of near-optimal solutions to choose from.

B. Network-Aware Service Composition

Several studies have dealt with service composition while considering the impact of QoS of the network. Authors in [3] present a network-based service composition technique for component services in large scale overlay networks. Similarly, another study in [4] introduces network awareness in composing domain services in multi-domain networks. The authors try to optimize delay and available bandwidth. However, these studies do not consider service composition in the context of services in the cloud. A recent approach [2] develop a service composition technique that minimizes network latency of composite services in the Cloud. The authors use a network model based on Euclidean distance technique to estimate latency of composite services. Their work is similar with this study in that they consider network latency. The main difference is that they only consider QoS of network while this work considers QoS of network alongside service QoS objectives.

C. Network Coordinate System

Network coordinate systems (NCS) are used to estimate latency between nodes in a network [2]. Their purpose is to reduce the delay observed from sending physical round trip time (RTT) packets between nodes across the network path. They operate by predicting RTT measurements for a fraction of nodes on the network path using techniques such Euclidean distance estimation (EDE) [5, 16, 17] and matrix factorization (MF) [21]. EDE embed network distances between nodes as metric spaces where known network distances (RTT) are mapped into a two dimensional Euclidean space in order to predict unknown network distances. EDE is however susceptible to triangle inequality [21] which leads to inaccurate estimates. MF on the other hand estimate unmeasured network distances by factorizing distance matrix consisting of both known and unknown RTT values using mathematical concepts such as gradient descent [19]. MF does not use metric spaces and so is resistant to triangle inequality and produces more accurate than EDE. Current EDE and MF models adopt a centralized approach towards RTT estimation. The approach usually involves using a central server to collectively predict RTT values for all the nodes in the network path. This means that if one RTT value is inaccurately estimated, then the accuracy of other RTT values could be negatively affected. In this work, the problem is avoided by adopting a novel decentralized MF approach within the network model where each BDS takes charge of predicting RTT with its neighbouring services independently of other services on the cloud network.

III. Proposed Approach

A. Problem Formulation

The problem can be described as follows:

Given a user request T that will require a set of tasks t_1 to t_n,

$$T = \{t_1, t_2, \ldots, t_n\},$$

Where *n* is the number of tasks to complete user request.

Each task is assigned a service group (*S*) which defines a set of candidate services (s_{ij}) capable of performing the given task (as seen in Fig. 4.),

$$S_i = \{s_{i1}, s_{i2}, \ldots, s_{ik_i}\}, \ \forall i \in [1..n], \forall j \in [1..k]$$

Where k_i is the number of candidate services in the *i*-th service group.

For each task, only one candidate service within its service group can be bound to the task t_i to form a composite service *C*.

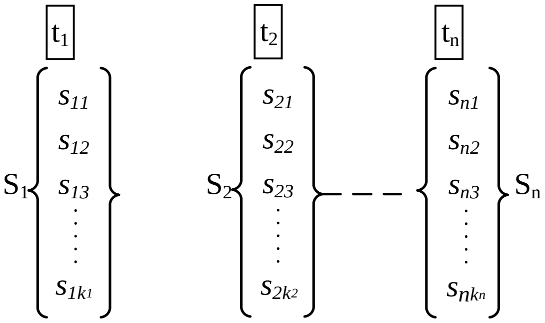


Fig. 4. Classification of candidate services into service groups and tasks

A composite service can be formed from the aggregation of candidate services;

$$C = \{s_{1j}, s_{2j}, \ldots, s_{nj}\}, \ \forall j \in [1..k], \ \forall i \in [1..n]$$

Where s_{ij} is the BDS bound to its service group S_i.

Also, given a set of QoS objectives (cost, execution time, response time and network latency) that need to be optimized, the end-to-end QoS value of a composite service ($Q(C)$) is calculated by combining individual QoS values of its services (one per task) based on the following expressions.

In order to determine end-to-end cost of composite service, cost for each service ($P(s_{ij})$) are combined

$$Q_P(C) = \sum_{i=1}^{n} P(s_{ij}) \tag{1}$$

Similarly, both end-to-end response time ($RT(s_{ij})$) and execution time ($ET(s_{ij})$) are aggregated respectively

$$Q_{RT}(C) = \sum_{i=1}^{n} RT(s_{ij}) \qquad (2),$$

$$Q_{ET}(C) = \sum_{i=1}^{n} ET(s_{ij}) \tag{3}$$

As for end-to-end network latency, RTT values are combined between each service in a given composite service

$$Q_{NL}(C) = \sum_{i=1}^{n} NL(s_{i,j}, s_{i+1,j}) \tag{4}$$

Where $NL(s_{i,j}, s_{i+1,j})$ represents the round trip time between each BDS in the cloud. Q_P, Q_{RT}, Q_{ET} and Q_{NL} represent end-to-end cost, response time, execution time and network latency of a composite service respectively.

Given weights w_P , w_{RT} , w_{ET} and w_{NL} which represent relative importance of QoS objectives from the user's perspective. Where,

$$\sum_{m=1}^{4} w_m = 1 \tag{5}$$

QoS objectives are normalized into fitness values using the expressions in Equations (6) and (7). Cost, response time and execution time are computed thus

$$F_m(C) = \sum_{\substack{i=1 \\ m\in\{P,RT,ET\}}}^{n} \left(\frac{Max_m(\mathrm{S}_i) - Q_m(\mathrm{s}_{ij})}{Max_m(\mathrm{S}_i) - Min_m(\mathrm{S}_i)} \right) \times w_m \qquad (6)$$

$$\forall j \in [1..k]$$

Network latency fitness value for composite service (F_{NL}) is determined by an expression in Equation (6) which normalizes the end-to-end network latency QoS (Q_{NL}).

$$F_{NL}(C) = \frac{Q_{NL}(C)}{H} \times w_{NL} \tag{7}$$

Where H is a constant which normalizes value of $Q_{NL}(C)$ in the range of [0 1].

The research problem becomes a constrained multi-objective optimization problem where the aim is to find a set of composite services with near-optimal fitness values,

$$C_{best} = Min[F(C)]$$

Subject to:

- Selection constraint: Only one candidate service can be selected per service group.
- Minimum ($q_m^{\min}$) and maximum ($q_m^{\max}$) QoS constraints:

$$\forall Q_p \in \left[q_p^{\min} \quad q_p^{\max} \right], \forall Q_{RT} \in \left[q_{RT}^{\min} \quad q_{RT}^{\max} \right],$$

$$\forall Q_{ET} \in \left[q_{ET}^{\min} \quad q_{ET}^{\max} \right]$$

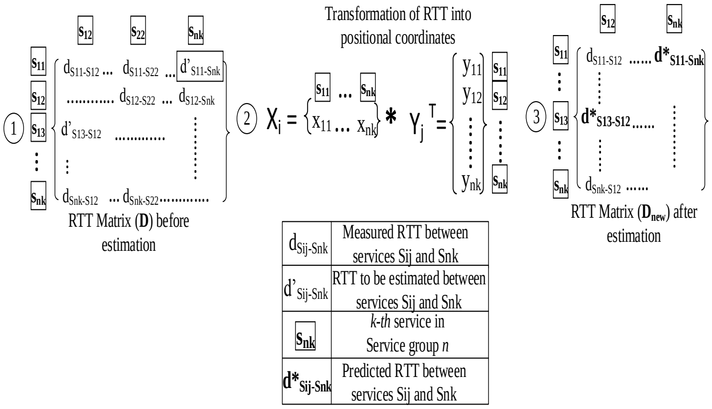
B. Network Model

In this study, a network model for estimating the RTT between BDS deployed on the cloud is adopted. The network model is composed of network coordinate system (NCS) based on Matrix factorization called LADMF (Learning-based Decentralized Matrix Factorization). Traditional MF techniques measure RTT ($d_{s_{ij}-s_{nk}}$) between a BDS and a subset of neighbors to build distance matrix *D*. These measurements are then used to predict RTT values ($d*_{s_{ij}-s_{nk}}$) for non-neighboring services as seen in Fig. 5. In mathematical terms, standard MF finds estimates of row matrix *X* and transposed column matrix *Y* that minimize the difference (ε) between measured RTT values in *D* and in estimated values in matrix D_{new},

$$\min[\varepsilon] \tag{8}$$

Where ε is the latency prediction error;

$$\varepsilon = (D - D_{new})^2 \tag{9}$$


Transformation of RTT into positional coordinates
1
2
3
RTT Matrix (D) before estimation
RTT Matrix (D_new) after estimation

$d_{Sij-Snk}$	Measured RTT between services Sij and Snk
$d'_{Sij-Snk}$	RTT to be estimated between services Sij and Snk
Snk	*k-th* service in Service group *n*
$d*_{Sij-Snk}$	Predicted RTT between services Sij and Snk

Fig. 5. Network distance estimation using matrix factorization

Also D_{new} is expressed as;

$$D_{new} = X * Y^T \tag{10}$$

Where X and Y are positional coordinates of all BDS on a given cloud network.

the standard MF technique is modified by adding learning automata concepts in order to further improve prediction accuracy of the estimation process. Instead of constructing a collective matrix (D_{new}) for all RTT estimates, LADMF decentralizes the process by allowing each BDS to estimate its own RTT values irrespective of other services. This is achieved by encoding each service as a learning automaton (LA) [5]. LA converts Equations (10) and (9) into Equations (11) and (12) respectively,

$$D_{ij\,new} = X_i * Y_j^T \tag{11}$$

$$\varepsilon = \left[\left(D_{ij} - D_{ij\,new} \right)^2 \right] \tag{12}$$

Where X_i is positional coordinate of *i*-th service, Y_j is positional coordinate of *j*-th neighbouring service, while D_{ij} .is the RTT between services *i* and *j*.

The effect is that each BDS will control their own path to RTT estimation without influencing estimation path of other services. Hence an inaccurate estimation of one service coordinate will not affect accuracy of other service coordinates.

In LADMF X_i and Y_j are encoded with additional LA parameters as seen in Fig. 6.

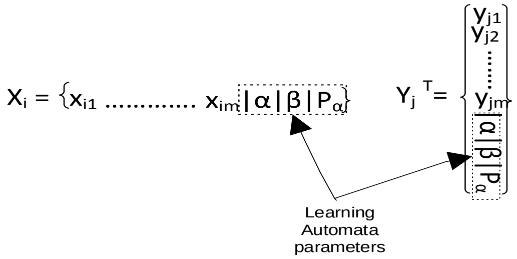

$X_i = \{x_{i1}$ .... $x_{im}|\alpha|\beta|P_\alpha\}$
$Y_j^T =$
y_{j1}
y_{j2}
y_{jm}
$\alpha|\beta|P_\alpha$
Learning Automata parameters

Fig. 6. Encoding of position vectors with LA parameters

Where

- α represents two alternative update strategies (α_1 and α_2) employed in updating position coordinates in X_i and Y_j:

$$\alpha \begin{cases} \alpha_1 \succ X_{i(new)} = D_{ij} Y_j (Y_j^T Y_j + (\Omega + J_1) I)^{-1}, \\ \quad Y_{j(new)} = D_{ij} X_i (X_i^T X_i + (\Omega + J_2) I)^{-1} \\ \\ \alpha_2 \succ X_{i(new)} = D_{ij} Y_j (Y_j^T Y_j + (\Omega - J_1) I)^{-1}, \\ \quad Y_{j(new)} = D_{ij} X_i (X_i^T X_i + (\Omega - J_2) I)^{-1} \end{cases} \tag{13}$$

Also

- Ω - Regularization parameter that controls speed of update
- J_1 and J_2 are constants
- I - Identity matrix
- β represents feedback for every action in α. β = {$\beta_{\alpha 1}$, $\beta_{\alpha 2}$}
- P_α is action probability which is determined from feedback of estimation error.

If feedback for action α_1 is good ($\beta_{\alpha 1}$ = 0 i.e. ε is improved) then action probability $P_{\alpha 1}$ is rewarded while $P_{\alpha 2}$ is penalized,

$$\beta_{\alpha_1} = 0 \begin{cases} P_{\alpha_1(new)} = P_{\alpha_1} + c(1 - P_{\alpha_1}) & c = 0.5, \\ & e = 0.005 * c \\ P_{\alpha_2(new)} = P_{\alpha_2} - e(1 - P_{\alpha_2}) & \end{cases} \tag{14}$$

Else if feedback is bad ($\beta_{\alpha 1}$ = 1 i.e. ε is not improved) then reverse is the case,

$$\beta_{\alpha_2} = 1 \begin{cases} P_{\alpha_2(new)} = P_{\alpha_2} + c(1 - P_{\alpha_2}) & c = 0.5, \\ & e = 0.005 * c \\ P_{\alpha_1(new)} = P_{\alpha_1} - e(1 - P_{\alpha_1}) & \end{cases} \tag{15}$$

Actions are evaluated and assigned probabilities based on error feedback which in this case is the estimation error (

$\min[\varepsilon]$). The action with the highest probability is selected as the next action. The process is continued until the estimation error is minimized. LADMF algorithm is outlined in Algorithm 1. Afterwards, estimated RTT values are aggregated to determine end-to-end network latency for a composite service via Equation (4).

Algorithm 1 LADMF Algorithm

```
Input:  D, max_iter, L,
Ouput:  Dnew
[X, Y] = function LADMF(D)
{       for(i =1: maxIter)
            for(j =1: max candidate service)
                $X \leftarrow$ rand(x)
                $Y \leftarrow$ rand(y)
                $\mathcal{E} \leftarrow w\,[D - (X * Y^T)]^2$
                if ( $\mathcal{E}$ is minimised)
                    Dnew $\leftarrow X * Y^T$
                    return
                endif
            endfor
        endfor
}
```

C. Network-Aware Service Composition Algorithm

A novel network-aware service composition technique based on non-dominated sort genetic algorithm (NSGA) is presented. When applying genetic algorithm to service composition problem, each genome represents a possible composite service and is encoded in form of array of numbers or genes, each gene in turn represents a task and can be assigned to any one of its candidate services (as seen in Fig. 7). State of the art NSGA initiates optimization process by building an initial generation of genomes then sorts individuals according to their fitness value and crowding distance. The best individuals are placed in a mating pool where they are altered by crossover and mutation operators to generate children that will populate subsequent generations. The whole process is repeated until optimization is reached. The state of the art NSGA algorithm is enhanced in order to be able to solve research problem. The improved algorithm called INSGA is described step by step as follows:

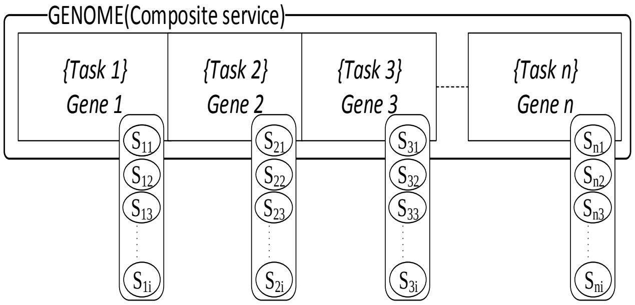


Fig. 7. Structure of composite service genome

Step.1. Initialization of Population. INSGA starts by randomly generating an initial population from the BDS that are part of the cloud. In order for this to be achieved, every service is first encoded as a two digit integer value. For example in Fig. 8, a BDS is encoded as "33" is the 3rd candidate service capable of executing task 3. In the next step only one candidate service is arbitrarily selected per task.

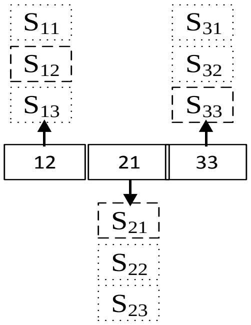


Fig. 8. Example of a composite service encoded as integer array

BDS QoS scores are then randomly initialized within their boundary constraints. With the aid of LADMF algorithm, the QoS scores are normalized and aggregated into values representative of composite service end-to-end cost, response time, execution time and network latency respectively.

Step 2. Ranking and Sorting. INSGA uses a non-dominated sorting technique that ranks individuals into different fronts according to the degree that they dominate other individuals in the population. A composite service C_i perfectly dominates another composite service C_j if all four fitness values of C_i are lower than the fitness values of C_j. Therefore C_i will be placed in a higher rank (front) than C_j. For each front, individuals are sorted in ascending order according to the magnitude of their fitness. This is used to establish the crowding distance (*CD*) which indicates the Euclidean distance between individual in the fitness value space. *CD* for a given composite service C_i is expressed as;

$$CD(C_i) = \frac{F(C_{i+1}) - F(C_{i-1})}{F_{\max} - F_{\min}} \qquad (16)$$

Where

- $CD(C_i)$ is the crowding distance for the *i*-th individual.
- $F(C_{i+1})$ represents the fitness value of individual succeeding *i*-th individual.
- $F(C_{i-1})$ represents the fitness value of individual preceding *i*-th individual
- $F_{\max}$ and $F_{\min}$ represent the maximum and minimum fitness values in population

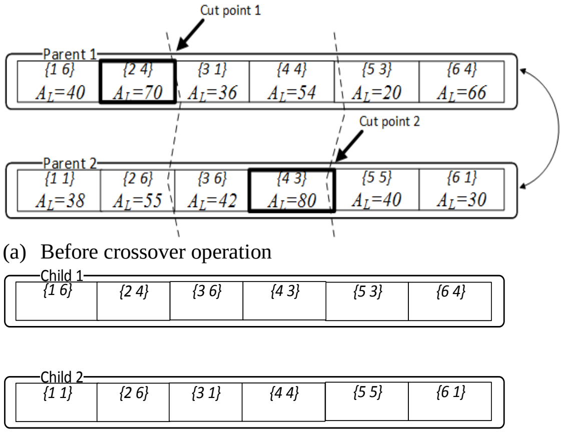
Step 3. Tournament Selection. A tournament selection of the best individuals that meet the user's satisfaction constraint is achieved to determine parents who will take part in crossover operation. The selection process ensures that only individuals with best fitness, rank and do not violate user constraint are selected for crossover operation.

Step 4. Crossover Operation. Crossover operation combines any two parents into offspring (children) that are quite different from their parents and can have superior properties of both parents. Traditional crossover operation picks arbitrary cut points where genes around cut points of one parent are replaced with genes of another parent to construct a set of children. INSGA employs a novel two-point crossover which cuts parents at two non-random cut points. The two cut points (one per parent) are chosen from points on each parent where average network latency is high. In order to determine which point on a parent constitutes poor average latency, every BDS assigned an average latency score (A_L) which is the arithmetic sum of RTT values over all outgoing paths divided by the number of outgoing paths from a given service,

$$A_L(s) = 1/|G| \sum_{\forall g \in G} Q_{NL}(g) \qquad (17)$$

Where A_L *(s)* represents average latency score in milliseconds (ms) for service *s*, G is number of outgoing paths from *s*, and Q_{NL} *(g)* is RTT value for a given path.

Once average latency scores are known, the crossover operator selects a cut point from each parent where A_L is maximum. After the cut points are known then the genes around those points are interchanged between both parents. This ensures that genes having highest A_L are interchanged with genes having lower A_L . Fig. 9 depicts how crossover operation is performed.



(a) Before crossover operation

(b) After crossover operation

Fig. 9. INSGA's two point crossover operation when cut point 1 and cut point 2 are not the same

When cut points 1 and 2 are the same for both parents then the crossover operation translates to a single point crossover. The impact of the crossover operator is that children produced are low latency versions of their parents as demonstrated by the results.

Step 5. Mutation Operation. The function of mutation operation is to adjust a parent into new offspring that closely resemble its parent with the aim of further improving parent fitness values and discourage trapping into local optima. The standard mutation operator adjusts parents by using a uniform distribution index (DI) [23]. DI controls degree of similarity between parents and their children. The value for DI influences the diversity of offsprings in the population. A new mutation operation is presented. The operator uses a variable distribution index whose value depends on a parent's crowding distance and fitness value for network latency. Each parent is going to be mutated according to the value of its distribution index which is computed using the following expression:

$$mum_{par\,i} = \left\{ \frac{H}{F_{NL}(par_i) + (1 - CD(par_i))} \right\} \times CD(par_i) \qquad (18)$$

Where

- mum_{par_i} is the distribution index for the parent.
- $F_{NL}(par_i)$ represents the parent's fitness value for network latency.
- $CD(par_i)$ indicates the parent's crowding distance.
- H is a constant.

The expression in Equation (18) will force a strong mutation for poor quality parents and a weak mutation for good quality parents. A large value for mum_{par_i} will indicate parent has good fitness and crowding distance therefore offspring's genes will closely resemble the parent (i.e. weak mutation), while a small value for mum_{par_i} indicates parent has poor fitness and crowding distance hence genes of offspring will differ greatly with the parent (i.e. strong mutation). This will ultimately improve the population diversity of new offspring and also increase the likelihood of finding the global solution. After mutation operation is performed, parents are replaced by newly formed off springs and the whole process is repeated until maximum number of generation is reached. INSGA algorithm is outlined in Algorithm 2 while the unique crossover and mutation operators are outlined in Algorithm 3 and 4 respectively.

Algorithm 2 INSGA Algorithm

```
Input: D, g, n, $\Omega$ , h, max_iter, no_states, state, actions_prob, rp_env, w, $J_1$, $J_2$,
Ouput: pop
Set environment parameters
pop ← Randomly generate population
$P$ ← Randomly generate QoS values of solutions
pop[$Q$ , $f$ ] ← Determine end-to-end QoS and fitness of solutions
pop ← Perform non-dominated sort (pop)
pop ← LADMF (Input)
While (gen ≠ maxgen)
  {
    pop ← tournament selection (pop)
    pop ← Crossover (pop)
    pop ← Perform non-dominated sort (pop)
    child_pop ← Mutation (pop)
    combination_pop ← pop + child_pop
    combination_pop ←Perform non dominated sort (combination_pop)
    pop ← replacement (combination_pop)
endWhile
}
```

Algorithm 3 INSGA Crossover operation

```
Input: pop
Ouput: Child
For(i = 1 to popsize)
{
Randomly pick Parent1 and Parent 2 from pop
Compute Average latency $A_L$ of Parent 1 and Parent 2
index1 ← Find cut point of Parent 1 with poorest latency
index2 ← Find cut point of Parent 2 with poorest latency
[Child 1, Child 2] ← Crossover genes for each parent around index 1 and index 2
[Child 1, Child 2] ← Determine end-to-end QoS and fitness of children
Child ←Add Child 1 and Child 2 in the child population.
endFor
}
```

Algorithm 4 INSGA Mutation operation

```
Input: pop
Ouput: Child
For(i = 1 to popsize)
  {
  Compute Dist. Index of pop(i) according to Equation (18)
   Child(i) ← Mutate genes of pop(i) according to DI
   Child(i) ← Determine end-to-end QoS and fitness of child
 endFor
 }
```

IV. Evaluation

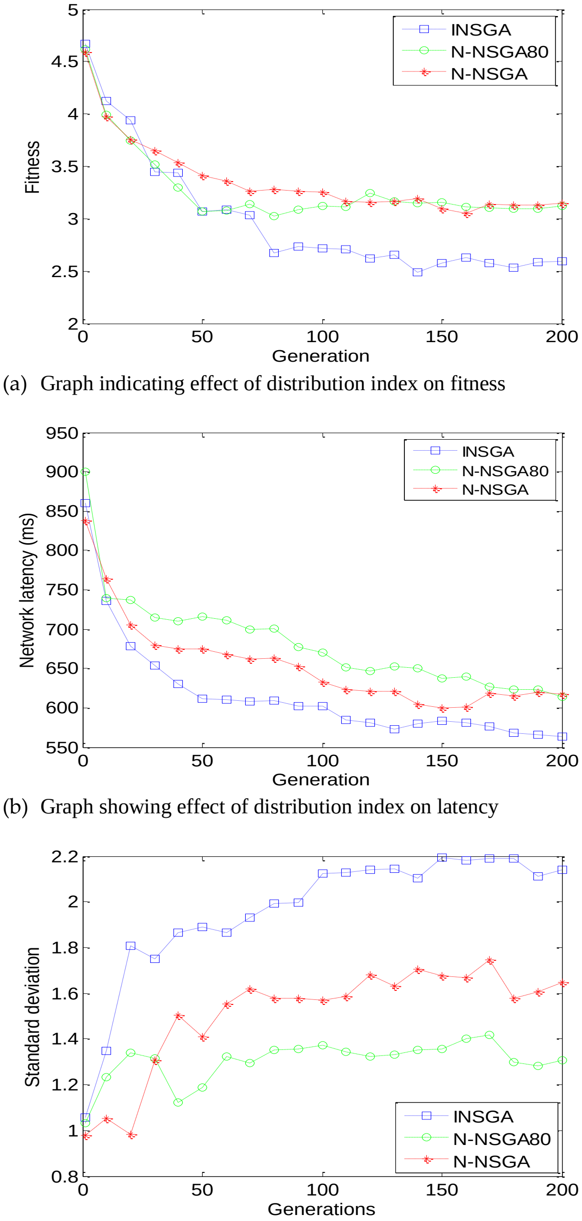
A. Setup

Experiments were run on a machine with Intel Core i7 CPU (3.8GHz) and with 8GB memory. All the algorithms and experiments are implemented in MATLAB 2013. A cloud network of BDS is simulated using planet lab meridian dataset [7] to provide RTT measurements between BDS. The dataset is chosen because it is expensive to implement a physically large cloud environment. The dataset contains symmetric round trip time (RTT) measurements between 1740 peer-to-peer nodes. Also, a test workflow is generated and will be used to evaluate INSGA algorithm. In the workflow, a set of thirteen tasks (t_1 to t_{13}) is defined. For each task, it is assume that each service group has equal number of candidate services for the sake of simplicity. The experiment is performed with 20 candidate services per task to simulate a large BDS cloud network.

B. Results and Discussion

To demonstrate the efficiency of INSGA, its fitness latency and population diversity are compared against other traditional algorithms such as Particle swarm optimization (PSO) [26] and Genetic algorithms N-NSGA [25] and S-NSGA [24] in different environmental situations such as variations in number of tasks, candidate services and distribution index. Given the probabilistic nature of the test algorithms, each algorithm is run 50 times to obtain average values for fitness, latency and standard deviation which is often used to measure diversity of population.

a) Impact of Distribution Index: In this experiment, an evaluation is done to determine the impact of distribution index on average fitness and population diversity of composite services. Here, the population size and maximum generation are set as 200 with network size of 260 services. In Fig. 10 (a) (b) and (c), it is observed that INSGA finds solutions with better fitness, latency diversity than N-NSGA and N-NSGA80. INSGA also avoids trapping in local optima while converging after 140 generations. This result shows that improvements in fitness, latency and population spread can be attributed to the proposed mutation and crossover operators.



(a) Graph indicating effect of distribution index on fitness

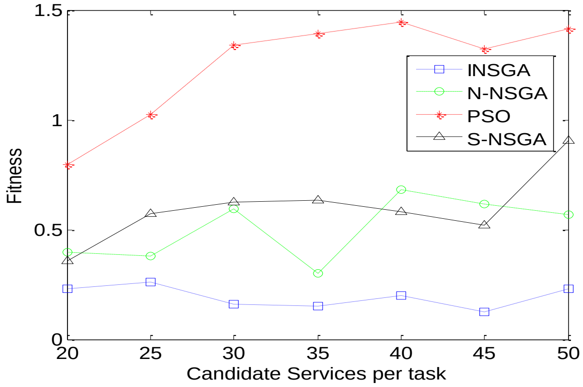
(b) Graph showing effect of distribution index on latency

(c) Graph showing effect of distribution index on population diversity

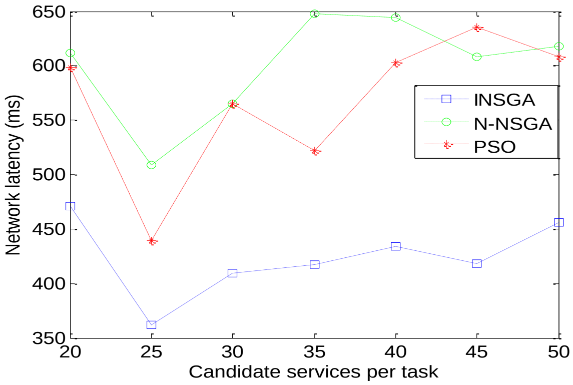
Fig. 10. Plot of Distribution index against fitness, latency and diversity of population

b) Size of Candidate Service per Task

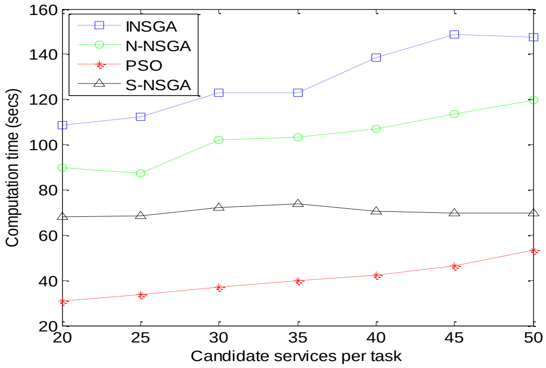
In this experiment, the number of candidate services per task is increased from 20 to 50 and evaluate the impact on network latency, fitness, computation time and standard deviation of population. In Fig. 11(a) and (b), it is noticed that an increase in size of candidate services may ultimately lead to better quality solutions for all test algorithms with the exception of PSO whose quality worsens. It can also be seen that INSGA finds solutions with the best fitness value and network latency when compared to the other algorithms. In Fig. 11(c) the computation times of all four algorithms are compared. It is observe that only INSGA has the highest computation time. This is as a result of the computational overhead generated by the network model. PSO has the lowest computation time which is about one third of INSGA's time. Also computation times for both N-NSGA, PSO and INSGA increase slightly with the number of candidate services except for S-NSGA whose computation time remains largely unchanged. Fig. 11(d) shows that increasing candidate service size doesn't influence the diversity of population for N-NSGA, PSO and S-NSGA. But in INSGA, standard deviation is improved slightly. Also, PSO shows the worst standard deviation while INSGA has the best standard deviation amongst the test algorithms.



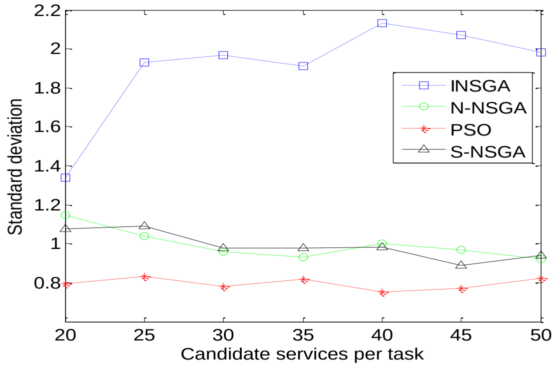
(a) Graph showing impact of number of candidate services on fitness



(b) Effect of number of candidate services on network latency



(c) Effect of number of candidate services per task on computation time

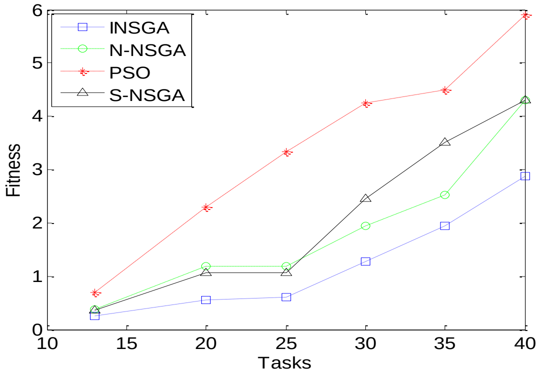


(d) Effect of number of candidate services per task on population diversity

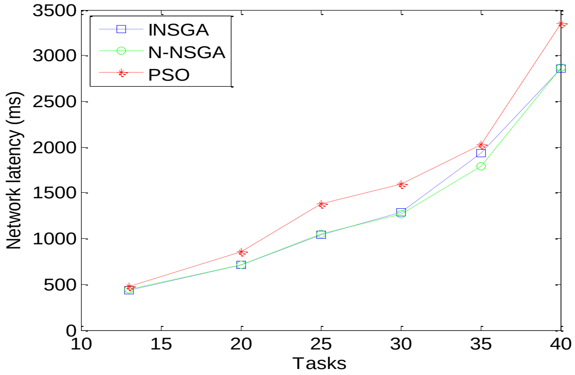
Fig. 11. Plot of candidate service size against fitness, network latency, computation time and standard deviation

c) Size of Tasks

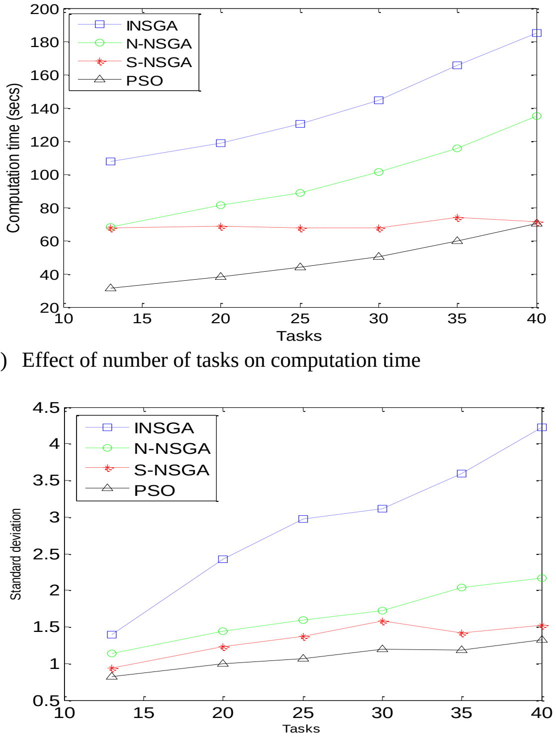
In this experiment the number of tasks are varied from 13 to 40 then the impact of fitness, network latency, computation time and standard deviation on the algorithms are determined. In Fig. 12 (a) and (b), it is observed that quality of fitness and network latency degrades with size of tasks for all test algorithms. INSGA is seen to produce the best quality solutions in terms of fitness and latency (tied with N-NSGA) while PSO produces worst quality of solutions. In Fig. 12 (c) a pattern similar to Fig. 11 (c) is observed, the only difference noticed is that computation time peaks at higher values when compared to graph in Fig. 12 (c). Lastly Fig. 12 (d) shows that population diversity increases linearly with size of tasks.



(a) Graph showing impact of number of tasks on fitness



(b) Graph showing impact of number of tasks on network latency



(c) Effect of number of tasks on computation time

(d) Effect of number of candidate services per task on standard deviation

Fig. 12. Plot of size of task against fitness, network latency, computation time and standard deviation

V. CONCLUSION

In this paper a novel approach to network-aware and QoS based service composition in the cloud is presented. Contrary to current works, this study separates QoS of network from service QoS. It consists of a network model which is composed of a novel network coordinate system called LADMF. LADMF uses matrix factorization to estimate the network latency (Round trip time) between BDS on the cloud. LADMF uses learning automata to encode service positional coordinates with additional learning parameters. This way the estimation process becomes decentralized where every service governs its own path to latency estimation. The latency information is then passed to a novel service composition algorithm based on non-dominated sort genetic algorithm called INSGA.

The aim of INSGA is to multi-objectively optimize cost, response time execution time and network latency QoS. INSGA uses a custom crossover and mutation operator. The crossover operator non-randomly picks two cut points where average latency is maximum while the mutation operator varies distribution index as a function of crowding distance and network latency. When compared with other state of the art service composition algorithms, results show that INSGA finds better quality solutions in terms of fitness, network latency and global search ability as indicated by its standard deviation.

REFERENCES

[1] Lifeng, Ai; "QoS-aware Web Service Composition Using Genetic Algorithms," PhD Thesis Queensland University of Technology, USA on, vol., no., pp.30, 2011.

[2] Adrian, K.; Fuyuki I.; Shinichi Honiden;,"Towards network-aware service composition in the cloud," In *Proceedings of the 21st international conference on World Wide Web* (WWW '12). ACM, New York, NY, USA, on, vol., no., pp.959-968, 2012.

[3] Jingwen Jin; Jin Liang; Jingyi Jin; Nahrstedt, K., "Large-Scale QoS-Aware Service-Oriented Networking with a Clustering-Based Approach,"*Computer Communications and Networks, 2007. ICCCN 2007. Proceedings of 16th International Conference on*, vol., no., pp.522, 528, 13-16 Aug. 2007.

[4] Jin Xiao; Boutaba, R., "QoS-aware service composition in large scale multi-domain networks," *Integrated Network Management, 2005. IM 2005. 2005 9th IFIP/IEEE International Symposium on*, vol., no., pp.397, 410, 15-19 May 2005.

[5] Frank Dabek; Russ Cox, frrans Kaashoek;"Vivaldi: A Decentralized Network Coordinate System," *ACM SIGCOMM '04 NY USA on*, vol., no., pp.15-26, 2004.

[6] Damien Saucez; "Securing Network Coordinate Systems," *in Master thesis: Université catholique de Louvain on,* vol.,no., pp.1-110, June 2007.

[7] Wong, B.; Slivkins, A.; Sirer, E.; "Meridian: A lightweight network location service without virtual coordinates," *In: Proc. the ACM SIGCOMM.*, vol., no., pp., 2005).

[8] Casati, F.; Georgakopoulos, D.;,Proceedings of the international workshop on Technologies for E-Services Roma, Italy on, vol., no., pp., September 2001.

[9] Tsur, S.; Abiteboul, S.; Agrawal, S.; Dayal, U., Klein, J; Weikum, G.;,"Are web Services the Next Revolution in e-commerce?," *Proceedings of the International Conference on very large databases on,* vol., no., pp. 614-617, September 2001.

[10] Bonet Blai;," Learning Depth-First Search: A Unified Approach to Heuristic Search in Deterministic and Non-Deterministic Settings, and its application to MDPs," *In Proceedings of ICAPS™, pp.3-23 2006.*

[11] Yan Gao; Jun Na; Bin Zhang; Lei Yang; Qiang Gong;, "Optimal Web Services Selection Using Dynamic Programming," *Computers and Communications, 2006. ISCC '06. Proceedings. 11th IEEE Symposium on*, vol., no., pp. 365- 370, 26-29 June 2006.

[12] HaiTao Song; Yanming Sun; Yingyu Yin; Shixiong Zheng;, "Dynamic Weaving of Security Aspects in Service Composition," *Service-Oriented System Engineering, 2006. SOSE '06. Second IEEE International Workshop*, vol., no., pp.189-196, Oct. 2006.

[13] Yoo, J.J.-W.; Kumara, S.; Dongwon Lee; Seog-Chan Oh; , "A Web Service Composition Framework Using Integer Programming with Non-functional Objectives and Constraints," *E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008 10th IEEE Conference on* , vol., no., pp.347-350, 21-24 July 2008.

[14] Zhifeng Gu; Bin Xu; Juanzi Li; "Inheritance-Aware Document-Driven Service Composition," E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2007. CEC/EEE 2007. The 9th IEEE International Conference on, vol., no., pp.513-516, 23-26 July 2007.

[15] Chen Ming; Wang Zhen wu;,"An Approach for Web Service Compositon Based on QoS and Discrete Particle Swarm Optimization," Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD 2007, Eight ACIS International Conference on, vol.2, no., pp. 37-41, August 2007.

[16] Adrian Klein; Fuyuki Ishikawa; Shinichi Honiden;,"Towards network-aware service composition in the cloud," In *Proceedings of the 21st international conference on World Wide Web* (WWW '12) on, vol., no., pp.959-968, 2012.

[17] Ng, T.S.E; Zhang, H.A.;'"A Network Positioning system for the Internet," *in Proc USENIX ATL, on*, vol., pp., 2004

[18] G. Canfora, M. D. Penta, R. Esposito, and M. L. Villani. An Approach for QoS-aware Service composition based on Genetic Algorithms. In GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation, pages 1069–1075, New York, NY, USA, 2005. ACM

[19] Yongjun Liao; Wei Du; Geurts, P.; Leduc, G., "DMFSGD: A Decentralized Matrix Factorization Algorithm for Network Distance

Prediction," *Networking, IEEE/ACM Transactions on* , vol.21, no.5, pp.1511,1524, Oct. 2013

[20] Rony Kay; “Pragmatic Network Latency Engineering Fundamental Facts and Analysis,” *cPacket Networks on* vol., no., pp.1-13, 2009

[21] Yongjun Liao; Geurts, P.; Leduc, G., "Network Distance Prediction Based on Decentralized Matrix Factorization," *Lecture Notes in Computer Science Springer* , vol.6091, no., pp.15-26, 2010

[22] Chen, S., Joshi, K., Hiltunen, M., Schlichting, R., W. Sanders; ”Link gradients: Predicting the impact of network latency,” on multi-tier applications. *In IEEE INFOCOM*, vol., no., pp., 2009.

[23] Mohammad Hamdan; “The Distribution Index of Polynomial Mutation for Evolutionary Multiobjective Optimisation Algorithms: An Experimental Study,” *Yarmouk University*,vol.,no.,pp.1-5,2012.

[24] L. Li; P. Yang; L. Ou; Z. Zhang; P.Cheng;”Genetic Algorithm-Based Multi-objective Optimisation from QoS-Aware Web Services Composition,” *In Springer Knowledge Science, Engineering and Management* vol.,6291,no.,pp.549-554, 2010.

[25] S. Umar; S. Ghazanfar; E. Gregory;.”Network Aware Composition for Internet of Thing Services,” *In Transactions on Networks and Communications*, vol.3, no.1, pp., 2015.

[26] H. Rezaie; N. NematBaksh; F.Mardukhi;”A Multi-Objective Particle Swarm Optimization for Web Service Composition” *In Springer Journal for Communications in Computer and Information Computer Science*, vol.88,no.,pp.112-122, 2010.

[27] M. Chen; T. Tan; J. Sun; Y. Liu; J. Pang; X. Li;”Verification of functional and Non-functional Requirements of Web Service Composition,” *Singapore University of Technology and Design*;vol., no., pp. 1-15, 2014.

[28] Jaeger, M.C.; Rojec-Goldmann, G.; Muhl, G., "QoS aggregation for Web service composition using workflow patterns," *Enterprise Distributed Object Computing Conference, 2004. EDOC 2004. Proceedings. Eighth IEEE International* , vol., no., pp.149,159, 20-24 Sept. 2004.

[29] X. Zhao; P. Huang; T. Liu; X. Li; "A Hybrid Clonal Selection Algorithm For Quality of service-aware Web Service Selection Problem." *International Journal Innovative Computing and Information Control IJICIC*, vol.8, no.12, pp. 8527-8544, 2012.

[30] Jennings, Roger. *Cloud Computing with the Windows Azure Platform*. John Wiley & Sons, vol., no., pp., 2010.

[31] H. Keqiang, A. Fisher, L. Wang, A. Gember, A. Akella, T.Ristenpart; "Next stop, the cloud: Understanding Modern Web Service Deployment in EC2 and Azure." *Proceedings of the 2013 conference on Internet measurement conference*. ACM, vol., no., pp.177-190, 2013.

[32] T. Magedanz, N. Blum, and S. Dutkowski; “Evolution of SOA concepts in telecommunications,” *IEEE Computer Magazine*, vol. 40, no. 11, pp. 46–50, 2007

SOHO: Information Security Awareness in the Aspect of Contingency Planning

Jason Maurer
College of Arts & Sciences
Regent University
Virginia Beach, Virginia U.S.A.

Brandon Clark
College of Arts & Sciences
Regent University
Virginia Beach, Virginia U.S.A.

Young B. Choi
College of Arts & Sciences
Regent University
Virginia Beach, Virginia U.S.A.

***Abstract*—This paper seeks to take general security awareness information for home and small business owners and make it understandable and accessible by looking at practical ways to keep valuable information accessible after an incident or disaster according to current methods. This paper will first review select general security awareness information, then take a look at some aspects of contingency planning and look at some basic practical techniques to use in order to protect systems and information from complete loss after an incident. Finally, the ground work for implementing an individualized plan for a small business office or home office will be laid and some practical steps to take will be recommended.**

Keywords—SOHO; Information Security; Contingency Planning; Small Office; Home Office

I. Introduction

It is recognized that cybersecurity is important but until recently has been neglected by many [1]. When a computer virus, power outage, loss of Internet access, vandalism, theft, hacker, fire, or some other disaster or incident occurs, will you be ready? When your computer shut down permanently or your hard drive malfunction, will you lose all your valuable information? If your filing cabinet and backed up files on the external hard drive in your desk drawer are under water, is all that information gone for good? If you anger an employee with access to your files stored in the cloud and they are all gone the next day, are you prepared to recover and continue with your business processes? These are just a few common scenarios that happen but are often ignored when it happens to someone else. Now is the time to prepare. Now is the time to plan. Our research seeks to inform you about an aspect of information security that is often avoided because of the time and resources it takes to do, with no obvious immediate results, and that is contingency planning. The goal of information system security is to maintain accessibility, confidentiality, and integrity [2] and information security in general encompasses of all your information, whether databases, files, or even printed papers and images, etc. Once you realize that you would be in a bad situation if you lost everything in your office, you need to know what to do and where to start. Our research is intended to get you thinking ahead and planning for the event that gets through all your layers of prevention and disrupts your operations and give a good foundation for you to start planning how to protect your information beginning with basic information security review.

II. Information Security Review

As first stated earlier, the goal of information security is to maintain accessibility, confidentiality, and integrity of all your information. This means electronic and physical information must remain at the determined level of privacy while also having accuracy and accessibility [2]. This is a very simple statement but a challenging goal. Information can be protected easily if it is not readily available and can be easily made available if it is not protected and can easily maintain accuracy if no one accesses it. Finding a good balance of these things if the goal of Information Security as shown in Figure 1 where you see examples of various imbalances and in the middle you see a balanced situation where only authorized users can access the information. Contingency planning is thinking ahead and preparing to maintain accessibility, confidentiality, and integrity of all your information so that you can continue operations with minimal down time in the event of an incident or disaster. Figure 2 shows the secure, accessible and accurate information being destroyed by incident or disaster. This triggers the recovery of the information using the contingency plan that was put in place for the specific type of event and restoring it to the original state of being secure, accessible and accurate. This begins with an assessment of the risks that your home or office might be susceptible to and then doing as much as possible to prevent or mitigate those risks and make a plan in case those preventions fail or don't work as expected.

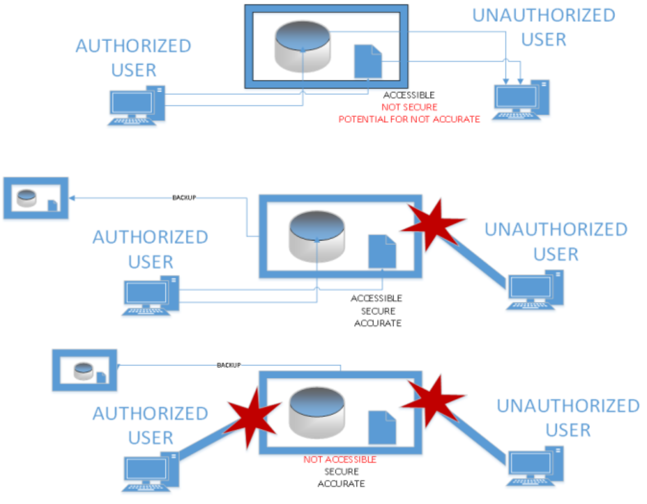


Fig. 1. Various Information Access Balances

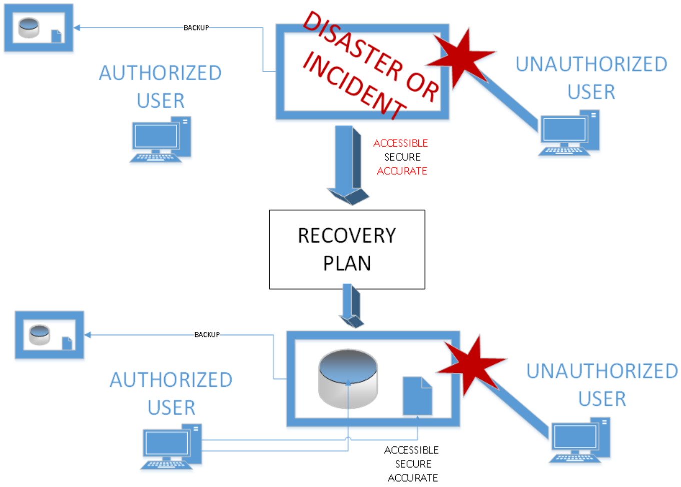


Fig. 2. Secure, Accessible, and Accurate Information Being Destroyed by Incident or Disaster

III. PLANNING

Planning is critical to maintaining a small office or home office for when incidents happen or disaster strikes. Having an insurance policy covering your home or business location is not enough. Contingency planning could be thought of as a form of self-insurance. It will cost time and resources to plan and prepare, as an insurance premium does, but when an incident happens, you will spend much less money to get things operational when it is all said and done because you have prepared and developed an appropriate plan. Contingency planning is not just about planning for major disasters such as fire, flood or Hurricane, but about recovering from situations that negatively affect operations such as suppliers going bankrupt, an important delivery being delayed, or even the entire office staff getting food poisoning at an office party [3]. "The goal of a resilient organization is to continue mission essential functions at all times during any type of disruption [4]."

A. *Why Plan?*

The security and continuity of your business processes is very important to any business or organization manager, especially in Small Office/Home Office (SOHO) setups. However, the authors speculate that most people either do not think that security is important, or think that it is second to everything else and often never get around to putting a plan in place. Putting planning in second place is flawed thinking as security and preparedness for incidents should be the first priority when setting up any information system. The threats to businesses and other organizations are growing and require great attention to information security technology to combat both people threats and the seeming increased occurrence of natural disasters [5].

Chris Stock, Director, Security Programs, TM Forum says, "I think there are still people out there who think, 'I don't have too much to worry about… They are just going to risk it and take the consequences if they are hacked [6]." More than 75% of the responders to a survey done for 100 small businesses in New York, New Jersey, and Connecticut after Hurricane Sandy did not have disaster recovery plans in place and "…on average, Staten Island small businesses lost more than $83,000 in revenues in the month that it took them to reopen. Annual revenues were reduced by approximately $200,000 to $500,000 following Sandy [7]." While this is just one specific incident of large scale, if those business had put the time and effort into creating a plan, they may have opened their doors much quicker and not lost nearly as much revenue. The damage from a disaster impacts more than the buildings, equipment, and processes. In certain situations, it also affects revenue, reputations, and can simply put people out of business.

B. *Why People Don't Plan*

As mentioned previously, the information published in Contingency Planning and Disaster Recovery by Cynthia Scarinci [7] uses a series of surveys from various sources to identify various shortfalls and issues that were experienced by accounting firms and small businesses in the aftermath of hurricane Sandy. In the article, the author shows where the responding business managers fell short and what hindered them from having plans in place. The top reason was that they simply did not know how. It can also be seen how much immediate losses were while recovering from Hurricane Sandy and that there was a general reduction in annual revenue after the disaster. There are two major issues to be addressed from the results of this survey. The first is that there are business owners that do not appreciate the importance of planning for the unexpected, often because of the time and cost commitment. Secondly, it is essential to educate these business owners on "how a plan can facilitate their recovery process [7]." This is exactly what our research seek to do, educate those people working in small offices and home offices and show how and where to start with contingency planning.

IV. TYPES OF PLANNING

There are several types of contingency plans that are beneficial for an organization to have as contingency planning can be broken into several different areas. Many of these are talked about in National Institute of Standards and Technology (NIST) publication SP 800-34 which covers the topic of contingency planning. Some of the plan types that apply to Contingency planning include Business Continuity Plan (BCP), Continuity of Operations (COOP) Plan, Disaster Recovery Plan (DRP), and the Information System Contingency Plan (ISCP). Each of these will be mentioned briefly to see how they fit into contingency planning. The majority of the information covered in this section is from the NIST SP 800-34 document.

A. *Business Continuity Plan*

The main purpose of creating a Business Continuity Plan (BCP) is to have information that will sustain the mission/business processes of an organization during and after a disruption. The examples in the NIST documentation given for these types of processes is the organization's payroll process or customer service process. This plan has information that keeps you going after a disaster. They say that a BCP "may be written for mission/business processes within a single business unit or may address the entire organization's processes" and the scope can be only priority functions [4].

B. Continuity of Operations Plan

Continuity of Operations (COOP) focuses on mission essential functions performed at an alternate site for up to 30 days. This typically includes items such as risk management, budgeting and acquisition of resources, order of succession, delegation of authority, vital records management, human capitol, and reconstitution. Our research does not go into depth on this type of planning as it is usually only used by organizations that are federally mandated to use them. Non-government organizations typically use a BCP described above [4].

C. Disaster Recovery Plan

Disaster recovery is just what it sounds like. It is used when an event happens and services are majorly disrupted, which are usually physical disruptions to a company's services that deny staff access to the primary facility and infrastructure for an extended period of time [4]. "A DRP is an information system-focused plan designed to restore operability of the target system, application, or computer facility infrastructure at an alternate site after an emergency [4]." There can be multiple contingency plans that all involve steps in recovery of many individual systems once an alternate location has been established [4]. Practically speaking, this is a pre-thought through plan of what you are going to do to restore your entire business if something happens that completely removes you from your current facility and disables many or all of your resources. It may seem overwhelming and impossible, and it is a large project but it can be done one step at a time.

D. Information System Contingency Plan

As mentioned above, the DRP can be made up of many contingency plans. Creating an Information System Contingency Plan (ISCP), established procedures to assess and recover a system following a disruption [4]. "The ISCP has key information needed for system recovery, including roles and responsibilities, inventory information, assessment procedures, detailed recovery procedures, and testing of a system [4]." One can distinguish this type of plan as being different from a DRP in that ISCP's can be implemented no matter the location. The same steps will be followed to restore a system where the DRP is about establishing an alternate site for operations [4].

E. Other Important Terms

NIST contingency planning documents also contains information on other important terms you will need to know and think about when working on your contingency plans. Maximum Tolerable Downtime (MTD) is the maximum acceptable amount of time a system can be down. This gives direction on recovery methods and detail of recovery procedures [4]. Critical operations with very low tolerable down time will require more resources and in depth planning to support a quick recovery time. Recovery Time Objective (RTO) "defines the maximum amount of time that a system resource can remain unavailable before there is an unacceptable impact on other system resources, supported mission/business processes, and the MTD." This is important for selecting the best methods suited to stay within the MTD [4]. Recovery Point Objective (RPO) basically determines how much data can be lost during the recovery. There is typically a point in history where the data can be restored from with the last backups. This determines the amount of data loss or how far back the restore point is allowed to be [4].

V. Starting to Plan

It is thought that the root cause of 80% of security incidents could be avoided by "doing the basics well [8]." To begin, you need to know what you do and how you do it. Start by developing a list of all the important or critical processes you have in place [3].

A. Processes

This can be a difficult task as everything from taking orders and delivering a product or processing payroll to purchasing paper for the copy machine has a standard process and can seem important. Distinguish between the processes that are critical to operations, payroll, payment processing, sales, customer support, etc. This is a good place to start. Take each one and list them out from most critical to the least critical.

B. Resources

Next determine what resources or systems are involved with each process. Use this information to determine the most important resources. If all your processes involve the Internet, then the Internet is going to be a critical resource along with making sure you have a working computer. "Examples of resources that should be identified include facilities, personnel, equipment, software, data files, system components, and vital records [4]."

C. Risks

Next, a manager will want to identify the risks applicable to their small businesses. This is done by doing a risk analysis. A Risk analysis is very involved if it is done comprehensively. You will "draw on detailed information such as project plans, financial data, security protocols, marketing forecasts, and other relevant information. However, it's an essential planning tool, and one that could save time, money, and reputations [9]." You need to determine what could harm or disrupt your business, such as file deletion, viruses, hackers, theft, hurricane, fire, etc. It is important to prioritize them or you will be overwhelmed. You don't want to start by making a plan for each threat. Start with those that will have the most impact and are the most likely to happen. At the same time you will be considering what you already have or will have in place to mitigate the risk. Looking at the threat of theft in a high crime area, the risk is high. If you mitigate the threat by installing a comprehensive top of the line security system along with armed security guards and 24/7 surveillance, the risk of theft is lower than the threat of a hacker if you have no firewall installed. "A good plan identifies all critical business functions, and it outlines ways to minimize losses [3]." The website for MindTools gives a list of well said guidelines for keep in mind while developing your plan. This list can be seen in the reference [3].

VI. Practical Steps

In this section, practical considerations of the planning process are discussed. Simple things like keeping a list of important contacts in a purse or wallet and making sure all employees have one also can make a huge difference on how contingency plans go. An alternate location, alternate internet access, as well as the cost versus benefit of everything will need to be considered. Backing up files is a huge part of recovering from disasters and certain incidents and will be emphasized a little more.

A. Alternate Internet Access

An area that is very standard for most business' staff to rely on is their Internet connection. It is a good idea to look into an alternate Internet option. It may be a hot spot that is compatible with cell towers or simply a different vendor that supplies an alternate site with service. Even if an alternate Internet Service Provider (ISP) is considered in the event that your standard provider's service ceases to operate, there is no guarantee that the alternate provider will not also be affected by the same conditions.

B. Alternate Location

Consider where you will go if your building is destroyed. Where will your new office location be until the current location is rebuilt or renovated? How far away will it be? If it is too close it may be affected by the same disaster, if it is too far you may not get all your employees there or be able to service the customers in your normal area. Do you need a hot site where all your data and systems are already on and running and fully staffed or do you need an empty room that you and a couple employees can go to for internet access on their laptops? These things need to be considered.

C. Weigh Cost versus Benefits

Throughout the process, when it comes time to make decisions, you will need to weigh the cost versus the benefit. This is where your impact analysis and risk assessment are valuable. If there is a critical process that your business cannot function without, then spare no expense to keep it running smooth. If a process is not needed to survive, do not invest a lot of time and money into having a plan B for it.

D. Virus Protection

Mitigate the threat of computer down time due to malware with virus protection. Many people do not know the benefit of having protection on each machine or do not know how or which solution to use. This is where a large company would have a large team of employees to address each question or issue. For small business or home offices those running the office will need to seek out a professional, or be educated.

E. Scan Paper Files

A simple method to backup information stored on paper, usually stored in file cabinets, is to scan in your papers and store them in searchable PDF format. It may take time to catch up, but integrating this step into your processes can save a lot of trouble in the future if a fire burns all the paper in your file cabinet. There are machines that can scan directly to a network storage device with no computer needed to scan the document making it simple and swift.

F. Backup Digital Data

Data is one thing that is very important in today's business world. It is important to have a plan in place in the event that a disaster, incident, or other event takes place. The thing is, many people and small businesses do not have any kind of backup solution [6]. It cannot be stressed enough that you must have multiple copies of your information backed up on different systems and even off site. Use either on an external server with RAID (Redundant Array of Independents Disks) options, on an external hard drive or the cloud to back up your data. Having multiple external hard drives that can be rotated is very beneficial. Even if you already backup your data in "the cloud" you may still want to have an offline copy of critical data that would cause severe problems and keep it in a location away from your office or home, even if it is just for peace of mind [10]. It is even wise to think about the situation of a disgruntled employee or even someone making a genuine mistake who deletes all the data out of the cloud or off a server. Being able to recover and restore data after an incident, whether a malware incident or a natural disaster, is critical to the continuity of a small office, whether at home or at a small business. The best policy is to have backups done at regular intervals to make sure the latest data is backed-up. An article by Santos and Bernadino called *Open Source Tools for Remote Incremental Backups on Linux* describes the easiest way to do a back-up is to copy an entire disk to a back-up disk [6]. This helps to prepare against drive failure. Luckily there are several free open-source utilities to help make back-ups more accessible [6]. The ones mentioned in the article are all Linux tools, but can be used to back-up other types of systems. The most popular tools include Rsync, Rdiff-backup, Duplicity, Areca Backup, and Link-Backup. The various tools are then tested to compare the performance of the various tools. After testing was complete, the authors determined that Rsync was the best if encryption and compression were not needed. If those where needed, Duplicity performed the best. However, Rsync was determined to be the most efficient tool to make simple back-ups [6]. There are also many cloud based vendors in existence that backup multiple versions of every file. They will even take a snapshot of an entire drive, operating system and all and can then restore it to that state if disaster strikes a machine.

VII. Conclusion

The topic of Information Security include many topics. In our research, several areas of information to get a small office started with contingency planning efforts for emergencies, incidents, and disasters were discussed. There will always be threats to the security of computer systems, but having a plan in place will help keep data secure, accurate, and accessible and keep a small office successful. Get a plan in place and maintain it, practice it and make it familiar to every employee so there are few questions during an incident or disaster.

References

[1] Fourie, L. et al. (2014). THE GLOBAL CYBER SECURITY WORKFORCE – AN ONGOING HUMAN CAPITOL CRISIS. Global Business and Technology Association.

[2] OIT Communications Group. (2014). definition-information-security. Retrieved 11 29, 2014, from oit.unlv.edu: https://oit.unlv.edu/network-and-security/definition-information-security

[3] Mind Tools Ltd. (n.d.). Contingency Planning. Retrieved 02 21, 2015, from mindtools.com: http://www.mindtools.com/pages/article/newLDR_51.htm

[4] Swanson, M. et al. (2010). Contingency Planning Guide for Federal Information Systems. SP 800-34. Retrieved from

http://csrc.nist.gov/publications/nistpubs/800-34-rev1/sp800-34-rev1_errata-Nov11-2010.pdf

[5] Williamson, J. (2014). Privacy & Security: Locking Value into the Digital Economy. TM Forum.

[6] Santos, A., & Bernardino, J. (2014). Open Source Tools for Remote Incremental Backups on Linux: An Experimental Evaluation. Journal Of Systems Integration (1804-2724), 5(3), 3-13

[7] Scarinci, C. A. (2014). Contingency Planning and Disaster Recovery after Hurricane Sandy. The CPA Journal, 60-63.

[8] Schreiner, S., Carpenter, M., Hamerstone, A., Coffey, C., Webb, N., & Rottinger, J. (2013). CyberOps Quick Start Guide: Human Factors, Version 1.2. Retrieved from TMForum: http://www.tmforum.org/GuideBooks/GB968CyberOpsQuick/50365/article.html

[9] Mind Tools. (n.d.). *Risk Analysis and Risk Management*. Retrieved 2015, from minddtools.com: http://www.mindtools.com/pages/article/newTMC_07.htm

[10] Brinson, L. C. (2014). backup cloud storage. Retrieved from how stuff works: http://computer.howstuffworks.com/backup-cloud-storage4.htm

An Improved Bees Algorithm for Real Parameter Optimization

Wasim A. Hussein[a,*], Shahnorbanun Sahran[b], Siti Norul Huda Sheikh Abdullah[c]
Pattern Recognition Research Group,
Center of Artificial Intelligence Technology,
Faculty of Information Systems and Technology,
Universiti Kebangsaan Malaysia,
43650 Bandar Baru Bangi, Malaysia

***Abstract*—The Bees Algorithm (BA) is a bee swarm-based search algorithm inspired by the foraging behavior of a swarm of honeybees. BA can be divided into four parts: the parameter tuning part, the initialization part, the local search part, and the global search part. Recently, BA based on Patch-Levy-based Initialization Algorithm (PLIA-BA) has been proposed. However, the initial stage remains an initial step, and its improvement is not enough for more challenging problem classes with different properties. The local and global search capabilities are also required to be enhanced to improve the quality of final solution and the convergence speed of PLIA-BA on such problems. Consequently, in this paper, a new local search algorithm has been adopted based on the Levy looping flights. Moreover, the mechanism of the global search has been enhanced to be closer to nature and based on the patch-Levy model adopted in the initialization algorithm (PLIA). The improvements in local and global search parts are incorporated into PLIA-BA to advise a new version of BA that is called Patch-Levy-based Bees Algorithm (PLBA). We investigate the performance of the proposed PLBA on a set of challenging benchmark functions. The results of the experiments indicate that PLBA significantly outperforms the other BA variants, including PLIA-BA and can produce comparable results with other state-of-the-art algorithms.**

Keywords*—Bees algorithm; Population initialization; Local search; Global search; Levy flight; Patch environment

I. Introduction

Most population-based metaheuristic algorithms, especially in recent years, are inspired by the collective intelligent behaviors of swarms of animals and insects such as fish, birds, bacteria, ants, termites, wasps, and fireflies. The biological studies showed that a swarm of such animals has impressive abilities to achieve fascinating, complex collective behaviors despite the simple behavior of each [1, 2]. It was found that the explanation of this amazing observation is the feature of self-organization that social animals have [2]. Self-organization can be considered as an organization without organizer in which no guidance from external or internal controller is needed [1]. Instead, decentralized control mechanisms are required for these social beings to update their activities by themselves based on some limited and local information [1]. These intelligent collective behaviors and the incredible capabilities of social animals to solve their daily life problems fascinated researchers to model their behaviors to solve real-world optimization problems. Then the model can be used as a base to develop artificial versions, either by tuning the model parameters using values outside the biological range or by assuming additional non-biological characteristics in the model design [2]. As a result, swarm intelligence in nature has been transferred from biological systems to artificial systems. Thus a new field called Swarm Intelligence (SI) was emerged under the field of Artificial Intelligence (AI), particularly under the Computational Intelligence (CI) field. Therefore, algorithms such as Ant Colony Optimization (ACO) [3], Particle Swarm Optimization (PSO) [4], Bacterial Foraging Optimization (BFO) [5], and the Firefly Algorithm (FA) [6] have been developed.

Among the social living beings that present interesting behavior and features are honeybees. The honeybees are very interesting creatures that exhibit several surprising intelligent behaviors when they behave as swarms of honeybees. Over the past decade, the collective intelligent behaviors of swarms of bees have been attracting the attention of researchers seeking to develop intelligent search algorithms. Examples of algorithms inspired by the behavior of bees include the Honey Bee Mating Optimization (HBMO) [7], the Artificial Bee Colony (ABC) algorithm [8], and Bee Colony Optimization (BCO) [9] algorithms.

One of the most recent bee-based algorithms is the Bees Algorithm (BA). BA is a population-based search algorithm proposed by Pham et al. [10] and inspired by the foraging behavior of swarms of honeybees searching for good food sources. Fundamentally, the algorithm performs a kind of exploitative local or neighborhood search combined with an exploratory global search. Both kinds of search modes implement uniform random search. In the global search, the scout bees are distributed uniformly at random to different areas of the search space to scout for potential solutions. In the local or neighborhood search, follower bees are recruited for patches found by scout bees to be more promising to exploit these patches. BA has been successfully applied to problems in many fields, such as function (continuous) optimization [10, 11], training neural networks [12], the job shop scheduling problem [11], and solving timetabling problems [13].

As a result of this, and also its simplicity and closeness to the actual behavior in nature, BA has garnered a significant amount of interest from researchers since its invention. We can divide BA into four parts or components: the parameter tuning or parameter setting part, the initialization part, the local search (exploitation) part, and the global search (exploration) part. Several studies have sought to improve BA and to enhance its performance by improving some of these parts. Some of these studies focused on the parameter tuning and setting part. The resulting improvements in these studies were gained by reducing the number of tunable parameters [12] or by developing tuning methods to tune the parameters of BA [14, 15]. Other studies focused on developing other concepts and strategies for the local (neighborhood) search part [16-18], or for both the local and global search parts [15, 19, 20].

However, limited attention has been paid to the improvement of the initialization part. In the initialization component, the foragers or searchers fly at random to initial resources. The initial location of foragers relative to the optimal resource (target) may affect the degree of optimality of other algorithm components. Therefore, recently, the initialization part of Basic BA has been paid attention and enhanced to improve the final solution quality and the convergence speed. Consequently, an initialization algorithm called the Patch-Levy-based Initialization Algorithm (PLIA) has been proposed and incorporated into Basic BA to adopt a BA version denoted by PLIA-BA [21]. However, the initial stage remains an initial step, and its improvement is not enough for more challenging problem classes with different properties. The local and global search capabilities are also required to be enhanced to improve the exploitativeness and exploration abilities of the algorithm, respectively. Thus, the quality of final solution and the convergence speed of PLIA-BA is improved on such problems. Hence, in this paper, the local and global search parts of PLIA-BA have been improved.

Most of the improvements achieved on BA were not inspired by natural bee behaviors. However, the imitation of the best characteristics in nature can lead to efficient metaheuristic algorithms [22]. Therefore, it is good to search for additional natural bee aspects that can be modeled in BA and improve its performance. There are numerous biological features in nature associated with honeybee foragers and food sources that can be beneficial if they are properly modeled and incorporated into Basic BA. Among these features that we can model are the distribution of food sources and the distribution of honeybees when they fly away from the hive foraging for food. In nature, flowers are usually distributed in patches that regenerate and are rarely completely depleted [23]. In addition, a scout honeybee flies away from the hive and moves randomly throughout the space according to Levy flight motion [24-26], which has been found to constitute the optimal search strategy [23, 24, 27]. During the harvesting season, a portion of the colony population is kept as scout bees foraging for new food sources on the global scale [10, 28]. Furthermore, in nature, it has been found that Levy looping search triggers the flight paths that is performed by honeybee foragers conducting a local search around a known food source [25]. Consequently, in this paper, we enhance PLIA-BA, and propose an improved version of BA called Patch-Levy-based Bees Algorithm (PLBA). PLBA utilizes the PLIA for initialization stage [21], a new local search algorithm that models Levy looping flights, and an enhanced global search that is improved based on the patch-Levy model adopted in PLIA. The proposed local search algorithm is called Greedy-Levy-based Local Search Algorithm (GLLSA).

Levy flights were first proposed as models of random walks in optimization algorithms by Gutowski [29], who advocated the use of Levy distributions instead of uniform and Gaussian distributions as mechanisms to generate the size of steps. The justification for this was that the frequent short steps generated by Levy distributions enable the optimization algorithm to intensify the search in regions around the current promising points. In addition, the occasional long jumps produced by Levy distributions help the optimization algorithm to escape from local optima. Levy flights were subsequently utilized as a search mechanism in many optimization algorithms, such as Levy Flights Optimization (LFO) algorithms [30], Cuckoo Search [31], the FA [32], the Bat algorithm [33], the Krill Herd (KH) algorithm [34], the ABC algorithm [35], and PLIA-BA [21]. However, to the best of the author’s knowledge, this is the first time Levy looping flights are being used to conduct the local search as in nature instead of freely roaming Levy flights.

The remainder of this paper is organized as follows. Section II reviews the Bees Algorithm based on Patch-Levy-based Initialization Algorithm (PLIA-BA). Section III describes the Levy flight and the proposed algorithm. Section IV presents the results of performance evaluations and experiments obtained for the proposed BA variant (PLBA) and compares these results with those obtained using other BA variants, including PLIA-BA and other state-of-the-art algorithms. Finally, section V concludes this paper.

II. The Bees Algorithm Based on Patch-Levy-Based Initialization Algorithm (PLIA-BA)

The PLIA-BA is an improved version over Basic BA incorporating the Patch-Levy-based initialization algorithm (PLIA) proposed by Hussein et al. [21] to enhance the initialization stage of Basic BA. The PLIA can be summarized as follows:

1) Divide the search space equally into P patches or segments.

2) Evaluate the vector of areas in the hive from which scout bees will fly ($c_1, c_2, \ldots, c_P$), represented by the centers of the patches (segments).

3) Divide and assign the (n) scout bees into the hive areas and evaluate the number of scout bees (nb) to be assigned in each area in the hive using the following equation:

$$nb = Int\left(\frac{n}{P}\right) \qquad (1)$$

4) Evaluate the number of remaining scout bees still not assigned (nrb) to be assigned in the last area in the hive using the following equation:

$$nrb = n\%P \qquad (2)$$

5) Set j = 1

Set Current area = c_j

While (Current area (c_j) is not the last area (c_P))

Distribute (*nb*) bees from the current area (c_j) inside the hive to the patches according to Levy flight distribution to constitute (*nb*) bees in the initial population

Set $j = j + 1$

Set Current area = c_j

End while

6) Distribute (nb+ nrb) bees from the last hive area to the patches according to Levy flight distribution to constitute (nb + nrb) bees in the initial population.

7) Return the constructed initial population of (n) scout bees.

III. Proposed Algorithm

In this paper, a new local search algorithm (GLLSA) that models the Levy lopping flights is developed and the global search is improved to be based on the patch-Levy model adopted in the initialization algorithm, PLIA. The initialization algorithm (PLIA) [21], the proposed local search (GLLSA), and the enhanced global search are incorporated into an enhanced version of BA called PLBA. In this section, Levy flights and the importance of the local and global search components are described and the proposed improvements in these components are presented. Then, the development of PLBA is outlined on the basis of the proposed improvements in local search and global search parts.

A. Levy flight

The existence of Levy flights as a movement pattern in biological organisms was first noted by Shlesinger and Klafter, who stated that the characteristics of Levy flights can be observed in foraging ants [36]. Levandowsky et al. subsequently introduced Levy walks as a swimming behavior in microorganisms [36]. Various empirical and theoretical studies have subsequently identified the Levy flight patterns and characteristics in the foraging of various animals and species such as the wandering albatross, reindeer, jackals, dinoflagellates, spider monkeys, sharks, bony fish, sea turtles, penguins, fruit flies (drosophila), bumblebees, and honeybees [36]. Evidence of using Levy flights as search patterns was also found in the foraging patterns of humans such as the Dobe Ju/'hoansi hunter-gatherers [37].

Levy flights are random walks that are named after Paul Levy, a French mathematician [29]. Levy flights consist of independent, randomly oriented steps with lengths *l*, drawn at random from an inverse power-law distribution with heavy and long tail, $p(l) \sim l^{-\mu}$ where $1 < \mu \le 3$. Levy flights are scale-free since they do not have any characteristic scale because of the divergent variance of $p(l)$, and they present the same fractal patterns regardless of the range over which they are viewed. The pattern in Levy flights can be described by many relatively short steps (corresponding to the detection range of the searcher) that are separated by occasional longer jumps.

Levy flights can be categorized into two categories: freely roaming Levy flights and Levy looping flights [25, 27]. The freely roaming flights consist of sequences of Levy steps, whereas the Levy looping flights comprise loops of the Levy steps. The freely roaming Levy flights constitute the optimal search strategy for the foragers searching for sparsely and randomly distributed patches. On the other hand, Levy looping search strategy is adopted by many foragers for exploiting a promising patch or area. In the freely roaming flights, the forager continues the search from the last location, whereas in the looping flights, the foragers restarts the search from the original promising location that represents the center of a potential patch until the target is found. This strategy aims at searching the local area of the potential patch more intensively. However, if no progress has been achieved with the time, the probability of finding the target in the vicinity of that location decreases, thus the original site is abandoned and freely roaming flights are adopted.

It can be easily observed from the definition of Levy flights that these flights constitute a series of displacements and orientations. Therefore, two steps are required to mimic Levy movements in nature and to implement these flights. The first step generates a random direction to mimic the random choice of direction by drawing it from a uniform random distribution. In the proposed algorithm, the direction is drawn from a uniform distribution between -1 and 1. The second step generates the step length that obeys a Levy distribution.

B. Local and Global Search Capabilities

The local search and global search components are two of the main components of the metaheuristic algorithms. These two components are equivalent to two important characteristics, which are the intensification and diversification characteristics, respectively [22]. In the intensification, the current promising areas are exploited and the best solution is selected. Whereas, in diversification, the search space is explored on the global scale. Therefore, the enhancement of the local and global search mechanisms of PLIA-BA can lead to significant improvement in the overall performance of PLIA-BA in terms of the solution quality, convergence speed, and success rate. In addition, as mentioned earlier, the imitation of the best characteristics in nature can lead to efficient metaheuristic algorithms [22]. Therefore, in this paper, the local search part of PLIA-BA is enhanced by proposing a new local search algorithm called Greedy Levy-based Local Search Algorithm (GLLSA) that models Levy looping search for the exploitation of patches. The global search part is also improved by employing the patch-Levy model adopted in the initialization stage.

1) Local Search

In this important stage, the areas (patches) of promising resources are exploited by recruiting other bees to these areas and conducting local search. Since there is a high chance to find the optimal solution near the good solutions, the exploitation step is considered an important step.

In Basic BA and PLIA-BA, the best sites (m) are selected for the local search. Then, the local search part is performed by first determining the size of the patch in which the search will be conducted and then the recruit bees are distributed uniformly at random to food sources inside the determined patch. However, in nature, it has been found that the honeybee foragers conducting local search around a known food source adopt Levy looping flights [25]. In the Levy looping search, the honeybee flies from the site whose local area is searched. If the target is found the foraging stops, otherwise the forager returns back to the same original site and randomly chooses a direction and length before foraging again. If no progress has been achieved with the time, the probability of finding the target in the vicinity of that location decreases, thus the original site is abandoned and freely roaming flights are adopted.

As a result of this natural behavior of bees conducting local search and since the imitation of the best characteristics in nature can lead to efficient algorithms, this paper proposes a new local search algorithm that very closely mimics the actual behavior of bees in nature. This local search algorithm is called the Greedy Levy-based Local Search Algorithm (GLLSA). In GLLSA, a recruit bee restarts the search from the current best site in a patch based on Levy flight distribution for a predetermined time until a better site is found. Once a better site is found, it becomes the current best site. The aim of this strategy is to search the area where the optimal solution is expected to be more intensively. In the proposed GLLSA algorithm, each recruit bee of the remaining recruit bees starts foraging from the last resource found to be the best source by previous recruit bees. This limitation is not found in nature but it is assumed to increase the property of being greedy of the BA and to increase the possibility of finding the optimal solution.

2) The Proposed Local Search Algorithm: Greedy Levy-based Local Search Algorithm (GLLSA)

Based on the concepts above, the Greedy Levy-based Local Search Algorithm (GLLSA) is proposed to conduct the local (neighborhood) search stage in the PLIA-BA. The pseudo-code of the algorithm is presented in Fig. 1.

Before conducting the local search, as in Basic BA and PLIA-BA, the fittest *m* sites of the fittest *m* bees are determined as the promising resources which require exploitation. Among these *m* sites, *e* sites need more exploitation by recruiting for them more bees than the remaining (*m* - *e*) sites. As in Basic BA and PLIA-BA, the number of recruited bees for the *e* sites is determined by the parameter *nep*. For the remaining (*m* - *e*) sites, the number of recruited bees is identified by the parameter *nsp*. After this selection process of sites for the local search, the local search algorithm starts.

As can be seen from the pseudo-code of the algorithm in Fig. 1, for each site (patch) of the *m* selected sites (patches), a number of bees are recruited (*RecruitBee*). *RecruitBee* can be either *nep* or *nsp* as mentioned before. Each site of the *m* sites represents the current best site in its patch. Every recruited bee achieves searching and foraging for better solutions or sites for some time (*t*) that can be found empirically by a trial and error and is usually simply set to *RecruitBee* or to one. During this time, the first recruit bee restarts searching from the same original best site as long as there is no better site has been found. If the recruit bee finds a better solution while it is searching for the predetermined time, this recruit bee stops searching, the site found is considered the current best site, and the next recruit bee starts the search. The foraging of the next recruit bee is continued by the same way from the last current best site. The same process is repeated for the remaining recruit bees. Finally, each original exploited site of the m sites is replaced with the last best solution found by the bees recruited for that site.

Each recruit bee conducts searching in a patch according to the following equation illustrated in the schematic diagram in Fig. 2:

$$b_i = s_{bestcur} + (2r-1)\times Levy\,(\gamma_2),\quad i = 1,2,\ldots,nep\,(\text{or } nsp) \qquad (3)$$

where b_i is the position of the i^{th} recruit bee, $s_{bestcur}$ is the current best site in the patch, $(2r-1)$ gives the direction of fly drawn from a uniform random distribution between -1 and 1, i.e. $(2r-1)\in \text{uniform}(-1,1)$, $r \in \text{uniform}\,(0,1)$ and $Levy\,(\gamma_2)$ represents the step length generated randomly from a Levy flight distribution with a search size or scale parameter γ_2.

This search size or scale is shrunk at each iteration of the algorithm after the local search, thus the step size of the bee generated from Levy distribution is decreased from time to time while foraging inside the neighborhood of the potential solution. The aim is to decrease the length of the long steps [30], thus, increasing the intensification and exploitativeness capability of the proposed PLBA and searching the region around the promising solution comprehensively. Additionally, the decrease of the long jumps can prevent the recruit bees from going beyond the regions of promising sites that may result from the dependency behavior of GLLSA where the foraging of each recruit bee depends on the foraging results of the previous one.

C. Global Search

In nature, during the harvesting season, the bee colony keeps a portion of the bee population as scout bees foraging for new food sources [10, 28]. These scout bees fly away from the hive and move throughout a patchily distributed environment at random according to Levy flight pattern [23, 26]. In the Basic BA, PLIA-BA and the proposed PLBA, this portion of scout bees is represented by *n* - *m* scout bees, thus a number (*n* - *m*) of scout bees is distributed to perform the global search.

In Basic BA and PLIA-BA, the global search is conducted by distributing this proportion of scout bees uniformly at random into the search space in the same way as in the initialization stage of Basic BA. On the other hand, the global search in the proposed PLBA is enhanced by imitating the natural behavior, which was the Levy motion in a patchily distributed environment. In the global search of the proposed PLBA, the scout bees are distributed from the hive areas which are chosen to be the same areas of the hive from which they are initially distributed. Then, these bees start to scout according to the Levy flight with search size or scale (γ_3) as in the initial step.

D. The Proposed Patch-Levy-based Bees Algorithm (PLBA)

Having presented the initialization algorithm (PLIA) in [21], the proposed local search algorithm (GLLSA) in Section 2), and the enhancement on global search in Section C, the main steps of the proposed Patch-Levy-based Bees Algorithm (PLBA) can be summarized as follows:

1. Initialize the population using the initialization algorithm (PLIA) [21].
2. Evaluate the fitness of the population.
3. While (stopping criterion not met)
   // Form new population
4. Select (*m*) sites for neighborhood search.
5. Conduct Local (neighborhood search) according to the proposed local search algorithm (GLLSA).
6. Redistribute the remaining bees (*n* - *m*) from their previous areas inside the hive to the patches according to Levy flight distribution to scout again and evaluate their fitness.
7. End while.

In the proposed PLBA, it is assumed that the number of patches in the initialization and global search parts are the same and represented by the same centers. However, in the local search, each selected site out of the m sites represents a patch center as in Basic BA and PLIA-BA.

IV. Experimental Setup and Results

A. Experimental Setup

1) Benchmark Functions

Most of the standard benchmark functions have some properties that could be exploited by some general-purpose metaheuristic algorithms to achieve good results [38]. Among these properties are the symmetric dimensions of the global optimum (i.e., the dimensions have the same values), and the location of the global optimum at origin, at the center of the search range, or on the bounds. For instance, if the optimization problem has symmetric dimensions, an optimization algorithm with a neighborhood operator that copies the value of one dimension to other dimensions can converge quickly to the global optimum. Such properties are considered problems in the standard test functions that should be avoided in order to test a novel global optimization algorithm [38].

As a result, to evaluate the proposed PLBA and compare it with other BA variants and with other state-of-the-art population-based algorithms, we employ a set of 25 challenging scalable benchmark functions with different characteristics and complexities provided for CEC'2005 session on real-parameter optimization [39]. These benchmark functions include functions with different properties such as unimodal, multimodal, shifted, rotated and unrotated, global minimum on the bounds and global minimum not on the bounds, with and without noise, separable and non-separable, hybrid composition functions, and so on. Experiments were carried out on the 10 and 30-dimensional versions of these challenging problems. Detailed information about these function optimization problems can be found in [39].

2) Performance Evaluation

In the evaluation of the algorithms on a set of test problems, especially the multimodal problems, some algorithms may have a small probability of success on a test function but converge fast. On the other hand, other algorithms may have a larger probability of success but converge slower [40]. Hence, it is good to evaluate the performance of an algorithm in terms of both convergence speed and success rate. One way to do this is to use the Success Performance (SP) measure [40]. SP is the expected number of function evaluations to achieve a certain success level (accuracy level) on a specific function. Lower value of SP for an algorithm on a problem means the algorithm is faster in solving that problem. SP can be defined as follows [39, 40]:

$$\mathrm{SP} = \frac{mean\left(\mathrm{FES}_{successful}\right)}{p_{success}}, \qquad (4)$$

where SP is the success performance of an algorithm on a specific function, $mean\left(\mathrm{FES}_{successful}\right)$ is the mean number of function evaluations for the successful runs, and $p_{success}$ is the probability of success of the algorithm on the function, and $p_{success} = \frac{RN_{successful}}{RN_{All}}$, where $RN_{successful}$ is the number of successful runs, and RN_{All} is the total number of runs performed. Each run of an algorithm on a function was deemed to be successful when the error value of that function is less than or equal to the acceptance (accuracy) level specified for that function.

From (4), it can be seen that $\mathrm{SP} = mean\left(\mathrm{FES}_{successful}\right)$ in the case of SR = 100% where $p_{success}$ = 1. Therefore, to compare the convergence speed of a set of algorithms with success rate (SR) of 100%, the mean number of evaluations can be used. However, when the success rates of the algorithms are of different values, the success rates should be taken into account in addition to the mean number of evaluations only in the successful runs in order to evaluate the convergence speed. Therefore, we use three performance metrics to evaluate the performance of the proposed PLBA algorithm; namely, the mean function error value, the success performance (SP), and the success rate (SR).

The mean function error value gives indication of the quality of the final solutions obtained by the proposed PLBA algorithm, whereas the success performance is employed to ascertain the convergence speed of this algorithm. The function error value is calculated as follows: $E = \left|f\left(x_{\mathrm{best}}\right) - f\left(x^*\right)\right|$, where x_{best} is the best solution and x^* is the global optimum. Success rate (SR) is one of the performance criteria used in the literature for evaluating the reliability of algorithms. It can be calculated as follows: $\mathrm{SR} = p_{success} \times 100$. In addition, convergence graphs for each function, in the case of *d* = 30, that plot the function error value against the number of function evaluations are used. These graphs show the median performance of the total runs. Further, the extra parameters relevant to the proposed PLBA are analyzed.

In this evaluation of the performance of PLBA, the 25 functions are employed using dimensions of *d* = 10, and *d* = 30 with maximum number of function evaluations (Max_FES) of 100,000, and 300,000 evaluations, respectively. The PLBA is run 25 times for each problem. Each run of the PLBA is terminated when the number of evaluations reaches the Max_FES or if the function error value is 10-8 or less. Each run of an algorithm is determined to be a successful when the error in the function value (*E*) is less than or equal to the accuracy level prescribed in [39] as follows: 1×10^{-6} for functions $f_1 - f_5$, 1×10^{-2} for functions $f_6 - f_{16}$, and 1×10^{-1} for functions $f_{17} - f_{25}$.

3) Parameter Settings

We compare the performance of PLBA with that of Basic BA, Shrinking-based BA, Standard BA, and PLIA-BA. Shrinking-based BA is an improved version of BA that utilizes a neighborhood shrinking procedure [41] and Standard BA is an improved variant that employs both shrinking and site abandonment procedures [42]. In addition, we conduct an additional set of experiments to compare the performance of PLBA with other population-based metaheuristic algorithms. These algorithms include Restart CMA-ES [43], DMS-PSO [44], SPC-PNX [45], DE [46], SaDE [47], ABC [48], and modified ABC (MABC) [48]. The Restart CMA-ES is the covariance matrix adaptation evolution strategy in which the population size is increased for each restart. DMS-PSO is a dynamic multi-swarm particle swarm optimization algorithm in which the swarms are regrouped frequently and the Quasi-Newton method is employed to improve the local search capability. SPC-PNX is a steady-state real parameter genetic algorithm with parent centric normal crossover. DE is the classical differential evolution algorithm. SaDE is a self-adaptive version of the differential evolution algorithm where the learning strategy and some control parameters are self-adapted by the previous learning experience. ABC is the classical artificial bee colony algorithm. Modified ABC (MABC) is a modified version of ABC in which two parameters were added. The first parameter was to determine the number of parameters to be mutated and perturbed and the second one was to specify the magnitude of this perturbation. The parameter settings of these algorithms can be found in their references.

In order to perform a fair comparison among the BAs, we execute the five versions of BA with the same setting for the common parameters: *n* = 20 for the number of scout bees, *m* = 3 for the number of selected sites, *e* = 1 for the number of elite sites, *nep* = 4 for the number of recruited bees for each site of the *e* sites, and *nsp* = 1 for the number of bees recruited for every site of the remaining (*m-e*) sites. In addition, the parameters relevant to each version are set to different values for different problems (See Appendix A: Tables A. 1 and A. 2). These parameters are closely related to the problem under study. It should be noted that it is good on these benchmarks to set small values for the Levy search size of local search (γ_2) to support the exploitation capability of the good regions and large values for the Levy search size of the global search (γ_3) to maintain the diversity of the population.

B. Experimental Results

The mean function error values achieved by PLBA after Max_FES for the 25 10-dimensional and 30-dimensional test problems are presented in Tables I and III, respectively. The success rate (SR) and success performance (SP) achieved by PLBA for the 25 functions in the case of d = 10 and d = 30 are tabulated in Tables II and IV.

From the results in Tables I and II, it could be observed that in the case of 10-dimensional problems, PLBA could find the global optimum for problems f_1, f_2, f_4, f_5, f_9, f_{12}, and f_{15} with success rate 100%, 100%, 100%, 100%, 100%, 8%, and 100%, respectively. In the case of 30-dimensional version of problems, PLBA could find the global optimum for problems f_1, f_2, f_7, f_9, and f_{15}, with success rate 100%, 96%, 68%, 100%, and 52%, respectively, as can be seen in Tables III and IV.

However, for problems $f_{16} - f_{25}$, PLBA, as other optimization algorithms as can be seen in Section E, could not find the global optimum for both 10 and 30-dimensional versions in all 25 runs owing to the high multimodality of these problems, the randomly located global optima and the huge number of randomly located deep local optima [38, 47].

C. Analysis of the Results of the Proposed PLBA

Among the unimodal functions $f_1 - f_5$, PLBA was able to solve the shifted sphere function (f_1), the Schwefel problem 1.2 with (f_2) and without (f_4) noise in fitness, and the Schwefel problem 2.6 with global optimum on bounds (f_5) in the case of the 10-dimensional problems in all 25 runs. Only the shifted rotated high conditioned elliptic function (f_3) was the problem that PLBA failed to optimize within the Max_FES. In the case of the 30-dimensional version, PLBA was able to optimize f_1 over all runs, and to solve f_2 most the time with success rate 96%. However, PLBA failed to solve the 30-dimensional version of f_4. The first three unimodal functions were of different number of conditions that make them of different complexities where f_3 is more difficult than f_2, and f_2 is more difficult than f_1. From the results, it could be observed that the high number of conditions of f_3 deteriorate the performance of the PLBA. On the other hand, although the noise disturbs the search in the optimization process, the noise had no significant effect on the performance of PLBA in the case of 10-dimensional problems. This can be confirmed by comparing the results of f_{16} (without noise) and f_{17} (f_{16} with noise) where the results of these two functions were almost the same. The same case with the best performing algorithm (Restart CMA-ES) where the noise affected its performance in the 30 dimensions case, whereas the noise had no significant effect on its performance in the case of *d* = 10, as can be seen in [43].

For the basic functions in the multimodal problem class, PLBA could find the global optimal solution of f_9, which was the shifted Rastrigin function in all 25 runs for both the 10 and 30-dimensional versions. It could be concluded from these results and from the results of PLIA-BA on the standard Rastrigin (f_6) [21] that shifting the global optimum of the Rastrigin function does not affect the performance of PLBA, which is an improved version of PLIA-BA. However, rotation seemed to pose difficulties to PLBA as can be inferred from the rotated version (f_{10}) of the function f_9. Additionally, PLBA was able to locate the global optimum of the shifted rotated Griewank function, f_7 for 30-dimensional version with success rate 68%. On the other hand, PLBA failed to solve this problem in the 10-dimensional case. This can be explained by that the increase in the maximum allowable number of function evaluations by factor of 3 (3.00E+05) gives PLBA the time to solve this function most the time despite the increase in the number of dimensions. This can also be observed with other algorithms on f_7 and other problems such as DE [46] on f_7, SaDE [47] on f_4 and f_7, DMS-PSO [44] on f_7, SPC-PNX [45] on f_6 and f_7, Restart CMA-ES [43] on f_3 and f_{15}, Shrinking-based BA on f_1, and Standard BA on f_1, f_2, and f_6 where the results of these algorithms on the stated functions in the 30-dimensional versions are better than those in the 10-dimensional versions. Furthermore, PLBA could optimize f_{12} with success rate 8% for the 10-dimensional problem.

However, the two multimodal expanded functions f_{13}, and f_{14}, and the eleven multimodal composition functions form the biggest challenge for the population-based metaheuristic algorithms as can be seen from the results. Nevertheless, PLBA was able to successfully optimize the hybrid composition problem f_{15} in all runs within the Max_FES in the 10-dimensional case and half the time (success rate 52%) in the case of 30-dimensional version.

An interesting note that all algorithms failed to solve the shifted rotated Ackley function with the global optimum on bounds, even though most of the algorithms such as the improved BAs [21] succeeded to optimize the standard Ackley function with success rate of 100%, as can be seen with f_9 in the benchmarks used in [21]. The different scaling employed in this advanced version of Ackley can accounted for this bad performance of the algorithms on the function [43]. The standard Ackley function has a flat area outside the search space ($[-32,32]^d$) and employing the linear transformation with 100 condition numbers to construct the challenging version (f_8) caused this flat area to be inside the search space [43]. Therefore, the search for the global optimum of this function looks like looking for a needle in a haystack [43, 44].

It should be noted that the different characteristics of the problems have a great effect on the performance of an optimization algorithm. However, the performance of an algorithm is not affected by these characteristics only, but also by the parameter tuning process that has a significant effect [45]. Therefore, fine-tuning of the parameters of PLBA, especially the extra parameters, on a specific problem may result in a good performance of the algorithm on that problem.

D. *Comparisons among the BAs*

By comparing the results of BA-based algorithms in Tables I - IV, it can be clearly seen that the performance of PLBA is much better than the other BA variants. In the case of 10-dimensional problems, all other BA variants failed to reach the success level within the predetermined Max_FES on all the test problems. However, in the case of 30-dimensional problems, Standard BA and Shrinking-based BA were able to optimize f_1 with success rate 100%, and f_2 with success rate 96% and 64%, respectively.

The convergence graphs of all BA-based algorithms including the proposed PLBA on some tested problems for d = 30 were presented in Figs. 3 and 4 to compare PLBA against other BA variants in terms of the convergence speed. It can be seen from these figures, that PLBA converges faster than other BA versions on most of the tested problems. Even though the PLBA was stuck in local optima, it could escape from that local optima after a number of evaluations. This can be accounted for by that the patch concept helps in spreading the solutions along the search space and creating diversity, then the frequent short steps of Levy flights cause rapid convergence and the rarely occurred long jumps help to increase the diversity and escape from local optima.

The good effect of controlling the combination of patch number and Levy search size can also be observed by investigating the convergence graphs of PLIA-BA where PLIA-BA performed better than Basic BA on almost all functions and equally or better than Shrinking-based BA or Standard BA on some functions. From Figs. 3 and 4, it can also be concluded that dynamic change in the step size of the local search was advantageous for BA in the Standard BA and Shrinking-based BA. However, the fast decrease in the step size might lead to premature convergence on the problems other than f_1, and f_2. The dynamic change in the search size or the scale of the Levy flight in the local search (γ_2) was also beneficial for the PLBA in many cases in controlling the length of the long steps [30]. In this case, there are still frequent short steps and rare long jumps and only the length of these steps are changed and reduced, especially the long ones. Thus, the exploitativeness of the PLBA in the local search area is increased with the possibility of escaping from local optima without making aggressive long jumps that can lead the search outside the good area.

We conducted statistical comparisons between PLBA and other BA variants in terms of the solution quality using the Friedman test. We conducted two sets of comparisons using d = 10 and d = 30 dimensions to evaluate the results and to check the behavior of the BA-based algorithms. Tables V and VI show the ranks achieved by Friedman test for both sets of comparisons. It can be clearly seen in these tables that PLBA ranked first in both sets of comparisons. Therefore, PLBA was the best performing algorithm in the 10 and 30 dimensions cases, whereas the worst one was Standard BA, followed by Basic BA in the case of d = 10, and Basic BA in the case of 30 dimensions. The *p*-values calculated using the statistic from the Friedman test were 0 and 0.000071 in the 10 and 30-dimensinal cases, respectively, as shown in Tables V and VI. These *p*-values suggested a highly significant difference among the performance of the BA-based algorithms considered.

Subsequently, we used two post hoc tests (Holm and Hochberg tests) [49] to compare PLBA with the rest of the BA-based algorithms. Tables VII and VIII show the adjusted *p*-

values obtained by the post hoc tests considering PLBA as the control method in the 10 and 30-dimensional cases, respectively. In the case of *d* = 10, PLBA showed a significant improvement over the Basic BA, Standard BA, and PLIA-BA at a level of significance α = 0.01, whereas no significant difference was found between Shrinking-based BA and PLBA. On the other hand, in the 30-dimensional case, PLBA showed a significant improvement over all other BA-based algorithms with a level of significance α = 0.01.

Although the Shrinking-based BA did not succeed in solving any function out of the 25 functions in the case of 10 dimensions, its results were generally good and no significant difference was detected between it and PLBA. It can also be observed that PLIA-BA outperformed Basic BA in both 10 and 30 dimensions cases and it also outperformed Standard BA in the case of *d* = 10. However, in both cases (*d* = 10 and *d* = 30) PLBA outperformed PLIA-BA and PLIA-BA could not successfully solve any problem. Thus, it can be concluded that although the improvement of the initialization algorithm of Basic BA can enhance the performance of Basic BA, it is not enough to solve more challenging problem classes.

E. Comparisons with Other State-of-the-art Algorithms

First, it should be pointed out that no results regarding the SR, and SP were tabulated in Tables IX and X for conventional ABC and MABC because their results have not been reported in [48]. From these tables, it can be observed that, all algorithms employed in these comparisons, including PLBA could locate the global optimum of f_1, and f_2 over all 25 trials in the 10-dimensional versions. Whereas in the 30-dimensional version, all algorithms were able to find the global optimum of f_1 in all runs, and of f_7 with different success rates. The most successful algorithm on this function f_7 was Restart CMA-ES with 100% success rate, followed by DMS-PSO, DE, SaDE, PLBA, and SPC-PNX with success rate 96%, 88%, 80%, 68%, and 64%, respectively. In addition, all algorithms failed to find the global optimal solution of the problems f_8, f_{13}, f_{14}, and $f_{16}-f_{25}$ because of the high multimodality of these problems [47].

An interesting finding that only PLBA was able to successfully solve the hybrid composition problem f_{15} in all 25 runs in the 10-dimensional version, as can be seen in Table IX. On the other hand, other algorithms either could not solve the problem in some runs such as DE, SaDE, and DMS-PSO or did not succeed in any run such as SPC-PNX and Restart CMA-ES. In addition, only PLBA was the algorithm that could find the global optimum for the 30-dimensional version of f_{15} where it achieved a success rate of 52%, as can be seen in Table X.

In general, it could be concluded from the comparisons that the shifted sphere function, f_1 is the simplest function in the test suite and the shifted rotated Ackley function (f_8), expanded functions (f_{13}, and f_{14}), and the composition functions ($f_{15}-f_{25}$) pose the greatest difficulty for the population-based optimization algorithms. In addition, from the comparisons between the 10 and 30-dimensional versions, it could be inferred that increasing the Max_FES and adjusting some parameters can significantly improve the results for an optimization algorithm for specific problems.

Regarding the convergence speed on the functions with non-zero success rate achieved by PLBA, the SP was used where the algorithm with smaller value is considered faster. It can be seen in Table IX that PLBA was faster than some algorithms and slower than others on f_1, f_5, f_9, and f_{15} in the case of *d* = 10. In this case, SaDE was the fastest on f_9 and f_{15}, followed by PLBA. On the other hand, PLBA was the slowest algorithm on f_2, f_4, and f_{12}. In the case of *d* = 30, as can be seen in Table X, PLBA was the fastest on f_{15} and after SaDE on f_9. For the other functions (f_1, f_2, and f_7) PLBA converged faster than some algorithms and slower than others.

To statistically analyze the results obtained by PLBA and other state-of-the-art metaheuristic algorithms, we also employed the Friedman test. The results of all 25 10-dimensional problems were reported for all algorithms considered in the comparisons. On the other hand, in the case of the 30-dimensional version, the results of SaDE [47], and DMS-PSO [44] for $f_{16}-f_{25}$, and $f_{20}-f_{25}$, respectively, were not reported in the literature. This means that some data are missing for some algorithms. In addition, there have not been results reported for the conventional ABC in the case of using d = 30 [48].

Therefore, we performed two sets of statistical comparisons. The first set was among all algorithms using the 10-dimensional version of all 25 benchmarks, whereas the second one was among all algorithms excluding ABC on the 30-dimensional version of the first 15 benchmarks. The ranks calculated through the Friedman test for the algorithms considered are tabulated in Tables XI and XII for both sets of comparisons. It can be clearly seen in these tables that Restart CMA-ES was the best performing algorithm in both cases of *d* = 10 and *d* = 30. Whereas the worst performing algorithm was SPC-PNX in the case of employing 10 dimensions, and was the MABC in the case of using 30 dimensions. For the proposed PLBA algorithm, it ranked sixth in the comparisons considering the 10-dimensional versions and ranked fourth in the comparisons employing the 30-dimensional versions. The *p*-values calculated using the statistic from the Friedman test were 0.010674 and 0.001348 in the 10 and 30-dimensinal cases, respectively, as shown in Tables XI and XII. These *p*-values suggested a significant difference among the performance of the algorithms considered.

Subsequently, we used Holm and Hochberg tests to test the specific difference between Restart CMA-ES and the rest of the algorithms. Tables XIII and XIV show the adjusted *p*-values obtained by Holm and Hochberg methods in the case of employing the 10 and 30–dimensional benchmarks, respectively, considering the Restart CMA-ES as the control method. The Holm and Hochberg methods suggested a significant difference in the performance between the Restart CMA-ES in one side and SPC-PNX, ABC, and PLBA in the other side in the case of using *d* = 10. It can be clearly seen in Table XIII that Restart CMA-ES demonstrated a highly significant improvement over SPC-PNX with a significance level of α = 0.01, and a significant improvement over ABC, and PLBA at a significance level of α = 0.1. However, in the case of the compression set on the 30-dimensional benchmarks, it can be clearly seen in Table XIV that Restart CMA-ES showed a significant improvement over SPC-PNX, DE, and MABC at a level of significance of α = 0.05. On the other hand, no significant difference was suggested by Holm and Hochberg between Restart CMA-ES and PLBA in the case of 30-dimensional benchmark problems. Thus, it can be concluded that PLBA perform well even if the number of dimensions increases.

V. Conclusion

Despite the importance of the initialization part, the initial stage remains an initial step and its improvement is not enough for more challenging problem classes with different properties. Thus, the local and global search capabilities were also enhanced to improve the quality of final solution and the convergence speed of PLIA-BA on such problems. In this paper, a new local search algorithm, GLLSA, which is based on the Levy looping flights, has been adopted. Moreover, the mechanism of the global search has been enhanced to be closer to nature and based on the patch-Levy model adopted in PLIA. Consequently, a new version of BA called PLBA, which utilizes the initialization algorithm (PLIA), local search algorithm (GLLSA), and the enhanced global search has been proposed

We investigated the performance of the proposed PLBA on a set of challenging benchmark functions that are free of the properties of the standard functions (e.g., symmetry) that can be exploited by the optimization methods. We compared the proposed PLBA with other state-of-the-art algorithms available in the literature. Additionally, we conducted comparisons between the BA variants.

The comparisons with BA-based algorithms have shown that PLBA significantly outperformed the other BA versions: Basic BA, Shrinking-based BA, Standard BA and PLIA-BA. The results validated what has been stated before that the improvement of the initialization stage is not enough for all problem types and the improvement of local and global search capabilities is required as well. The experiments have also indicated that PLBA was able to produce comparable results with other state-of-the-art algorithms on the 10 and 30-dimensional challenging problems employed in the comparisons.

The problems employed in this work were static problems. Future work will focus on evaluating and validating the performance of the proposed PLBA on a set of recently proposed dynamic optimization problems.

Acknowledgment

The authors would like to thank the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, for providing facilities and financial support under Fundamental Research Grant Scheme No. AP-2102-019 entitled “Automated Medical Imaging Diagnostic Based on Four Critical Diseases: Brain, Breast, Prostate and Lung Cancer”.

References

[1] S. Garnier, J. Gautrais, G. Theraulaz, The biological principles of swarm intelligence, Swarm Intelligence, 1 (2007) 3-31.

[2] E. Bonabeau, M. Dorigo, G. Theraulaz, Swarm intelligence: from natural to artificial systems, Oxford university press, 1999.

[3] M. Dorigo, C. Blum, Ant colony optimization theory: A survey, Theoretical computer science, 344 (2005) 243-278.

[4] R. Poli, J. Kennedy, T. Blackwell, Particle swarm optimization, Swarm Intelligence, 1 (2007) 33-57.

[5] K.M. Passino, Biomimicry of bacterial foraging for distributed optimization and control, in: IEEE Control Systems 2002, pp. 52-67.

[6] X.-S. Yang, Firefly algorithms for multimodal optimization, in: O. Watanabe, T. Zeugmann (Eds.) Stochastic algorithms: foundations and applications, Springer, Berlin, Heidelberg, 2009, pp. 169-178.

[7] H.A. Abbass, MBO: Marriage in honey bees optimization-A haplometrosis polygynous swarming approach, in: Proceedings of the 2001 Congress on Evolutionary Computation, IEEE, Seoul, 2001, pp. 207-214.

[8] D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, Journal of global optimization, 39 (2007) 459-471.

[9] D. Teodorović, M. Dell’Orco, Bee colony optimization–a cooperative learning approach to complex transportation problems, in: Advanced OR and AI Methods in Transportation: Proceedings of 10th Meeting of EURO Working Group on Transportation Poznan, Poland, 2005, pp. 51-60.

[10] D. Pham, A. Ghanbarzadeh, E. Koc, S. Otri, S. Rahim, M. Zaidi, The bees algorithm-a novel tool for complex optimisation problems, in: Proceedings of the 2nd Virtual International Conference on Intelligent Production Machines and Systems (IPROMS 2006), Elsevier Science Ltd, Cardiff, UK, 2006, pp. 454-459.

[11] B. Yuce, D. Pham, M. Packianather, E. Mastrocinque, An enhancement to the Bees Algorithm with slope angle computation and Hill Climbing Algorithm and its applications on scheduling and continuous-type optimisation problem, Production & Manufacturing Research, 3 (2015) 3-19.

[12] D. Pham, H.A. Darwish, Using the bees algorithm with Kalman filtering to train an artificial neural network for pattern classification, Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 224 (2010) 885-892.

[13] S. Abdullah, M. Alzaqebah, A hybrid self-adaptive bees algorithm for examination timetabling problems, Applied Soft Computing, 13 (2013) 3608-3620.

[14] S. Otri, Improving the bees algorithm for complex optimisation problems, in, Cardiff University, 2011.

[15] Q. Pham, D. Pham, M. Castellani, A modified bees algorithm and a statistics-based method for tuning its parameters, Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 226 (2012) 287-301.

[16] M. Packianather, M. Landy, D. Pham, Enhancing the speed of the Bees Algorithm using Pheromone-based Recruitment, in: 7th IEEE International Conference on Industrial Informatics (INDIN 2009) IEEE, Cardiff, Wales, 2009, pp. 789-794.

[17] S.A. Ahmad, D.T. Pham, K.W. Ng, M.C. Ang, TRIZ-inspired Asymmetrical Search Neighborhood in the Bees Algorithm, in: Sixth Asia Modelling Symposium (AMS), IEEE, Bali, 2012, pp. 29-33.

[18] B. Yuce, M.S. Packianather, E. Mastrocinque, D.T. Pham, A. Lambiase, Honey Bees Inspired Optimization Method: The Bees Algorithm, Insects, 4 (2013) 646-662.

[19] A. Ghanbarzadeh, Bees Algorithm: a novel optimisation tool, in, Cardiff University, 2007.

[20] N. Shatnawi, S. Sahran, M. Faidzul, A Memory-based Bees Algorithm: An Enhancement, Journal of Applied Sciences, 13 (2013) 497-502.

[21] W.A. Hussein, S. Sahran, S.N.H. Sheikh Abdullah, Patch-Levy-based initialization algorithm for Bees Algorithm, Applied Soft Computing, 23 (2014) 104-121.

[22] X.-S. Yang, Review of meta-heuristics and generalised evolutionary walk algorithm, International Journal of Bio-Inspired Computation, 3 (2011) 77-84.

[24] A. Reynolds, Cooperative random Lévy flight searches and the flight patterns of honeybees, Physics letters A, 354 (2006) 384-388.

[25] A.M. Reynolds, A.D. Smith, D.R. Reynolds, N.L. Carreck, J.L. Osborne, Honeybees perform optimal scale-free searching flights when attempting to locate a food source, Journal of Experimental Biology, 210 (2007) 3763-3770.

[26] P. Bailis, R. Nagpal, J. Werfel, Positional communication and private information in honeybee foraging models, in: Proceedings of the 7th international conference on Swarm Intelligence, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 263-274.

[27] A.M. Reynolds, A.D. Smith, R. Menzel, U. Greggers, D.R. Reynolds, J.R. Riley, Displaced honey bees perform optimal scale-free search flights, Ecology, 88 (2007) 1955-1961.

[28] T.D. Seeley, The wisdom of the hive: the social physiology of honey bee colonies, Harvard University Press, Cambridge, Massachusetts, 1995.

[29] M. Gutowski, Lévy flights as an underlying mechanism for global optimization algorithms, arXiv preprint math-ph/0106003, (2001).

[30] T. Tran, T.T. Nguyen, H.L. Nguyen, Global optimization using L'evy flights, in: Proceedings of the 3rd National Symposium on Research, Development and Application of Information and Communication Technology (ICT.rda), Hanoi, Vietnam, 2004.

[31] X.-S. Yang, S. Deb, Cuckoo search via Lévy flights, in: World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), IEEE, Coimbatore, 2009, pp. 210-214.

[32] X.-S. Yang, Firefly algorithm, Levy flights and global optimization, in: M. Bramer, R. Ellis, M. Petridis (Eds.) Research and Development in Intelligent Systems XXVI, Springer, London, 2010, pp. 209-218.

[33] J. Xie, Y. Zhou, H. Chen, A Novel Bat Algorithm Based on Differential Operator and Lévy Flights Trajectory, Computational intelligence and neuroscience, 2013 (2013) 13.

[34] G. Wang, L. Guo, A.H. Gandomi, L. Cao, A.H. Alavi, H. Duan, J. Li, Lévy-Flight Krill Herd Algorithm, Mathematical Problems in Engineering, 2013 (2013) 14.

[35] H. Sharma, J.C. Bansal, K. Arya, Opposition based lévy flight artificial bee colony, Memetic Computing, 5 (2013) 213-227.

[36] G. Viswanathan, E. Raposo, M. Da Luz, Lévy flights and superdiffusion in the context of biological encounters and random searches, Physics of Life Reviews, 5 (2008) 133-150.

[37] C.T. Brown, L.S. Liebovitch, R. Glendon, Lévy flights in Dobe Ju/'hoansi foraging patterns, Human Ecology, 35 (2007) 129-138.

[23] G. Viswanathan, S.V. Buldyrev, S. Havlin, M. Da Luz, E. Raposo, H.E. Stanley, Optimizing the success of random searches, Nature, 401 (1999) 911-914.

[38] J. Liang, P. Suganthan, K. Deb, Novel composition test functions for numerical global optimization, in: Proceedings of the 2005 IEEE on Swarm Intelligence Symposium (SIS 2005) IEEE, 2005, pp. 68-75.

[39] P.N. Suganthan, N. Hansen, J.J. Liang, K. Deb, Y.-P. Chen, A. Auger, S. Tiwari, Problem definitions and evaluation criteria for the CEC 2005 special session on real-parameter optimization, in, Nanyang Technological University, Singapore and KanGAL, 2005.

[40] A. Auger, N. Hansen, Performance evaluation of an advanced local search evolutionary algorithm, in: The 2005 IEEE Congress on Evolutionary Computation IEEE, 2005, pp. 1777-1784.

[41] D. Pham, E. Koç, Design of a two-dimensional recursive filter using the bees algorithm, International Journal of Automation and Computing, 7 (2010) 399-402.

[42] D. Pham, M. Castellani, Benchmarking and comparison of nature-inspired population-based continuous optimisation algorithms, Soft Computing, 18 (2014) 871-903.

[43] A. Auger, N. Hansen, A restart CMA evolution strategy with increasing population size, in: The 2005 IEEE Congress on Evolutionary Computation IEEE, 2005, pp. 1769-1776.

[44] J.J. Liang, P.N. Suganthan, Dynamic multi-swarm particle swarm optimizer with local search, in: The 2005 IEEE Congress on Evolutionary Computation, Ieee, Edinburgh, Scotland, 2005, pp. 522-528.

[45] P.J. Ballester, J. Stephenson, J.N. Carter, K. Gallagher, Real-parameter optimization performance study on the CEC-2005 benchmark with SPC-PNX, in: The 2005 IEEE Congress on Evolutionary Computation, IEEE, Edinburgh, Scotland, 2005, pp. 498-505.

[46] J. Ronkkonen, S. Kukkonen, K.V. Price, Real-parameter optimization with differential evolution, in: The 2005 IEEE Congress on Evolutionary Computation, IEEE, Edinburgh, Scotland, 2005, pp. 506-513.

[47] A.K. Qin, P.N. Suganthan, Self-adaptive differential evolution algorithm for numerical optimization, in: The 2005 IEEE Congress on Evolutionary Computation, IEEE, 2005, pp. 1785-1791.

[48] B. Akay, D. Karaboga, A modified artificial bee colony algorithm for real-parameter optimization, Information Sciences, 192 (2012) 120-142.

[49] J. Derrac, S. García, D. Molina, F. Herrera, A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms, Swarm and Evolutionary Computation, 1 (2011) 3-18.

TABLE I. MEAN ERROR VALUES ACHIEVED FOR PROBLEMS $f_1 - f_{25}$ (d = 10) BY BAS

Problem	PLBA		PLIA-BA		Standard BA		Shrinking-based BA		Basic BA	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
f_1	**0.00E+00**	**0.00E+00**	4.78E-03	1.05E-03	5.40E-02	1.03E-01	1.43E-04	3.30E-05	4.77E-03	9.54E-04
f_2	**3.98E-13**	**2.83E-13**	5.52E-03	1.74E-03	5.28E-03	1.63E-02	1.93E-04	6.66E-05	5.60E-03	1.46E-03
f_3	**8.87E+03**	**1.74E+04**	1.53E+05	1.16E+05	8.63E+04	5.08E+04	1.37E+05	1.30E+05	1.21E+05	1.15E+05
f_4	**2.38E-09**	**3.23E-09**	6.84E-01	1.81E-01	5.63E-02	2.04E-01	7.31E-03	1.77E-03	6.95E-01	1.63E-01
f_5	**6.08E-11**	**7.48E-11**	1.56E+01	4.98E+00	6.65E+01	2.43E+02	6.51E+01	2.38E+02	1.65E+01	3.34E+00
f_6	**1.25E+01**	**2.19E+01**	1.66E+02	4.44E+02	1.44E+03	2.65E+03	1.76E+01	5.01E+01	6.16E+01	1.06E+02
f_7	**2.13E-01**	**8.42E-02**	6.65E-01	1.13E-01	7.94E-01	1.72E-01	7.57E-01	1.28E-01	7.64E-01	1.26E-01
f_8	**2.00E+01**	4.80E-02	**2.00E+01**	**2.31E-02**	2.04E+01	6.06E-02	**2.00E+01**	2.41E-02	**2.00E+01**	3.85E-02
f_9	**0.00E+00**	**0.00E+00**	2.80E+01	7.66E+00	3.08E+01	1.62E+01	2.69E+01	9.19E+00	2.75E+01	9.45E+00
f_{10}	2.16E+01	1.04E+01	3.64E+01	7.96E+00	3.48E+01	2.05E+01	**1.94E+01**	**8.72E+00**	3.64E+01	4.67E+00
f_{11}	4.23E+00	1.38E+00	5.33E+00	1.17E+00	9.19E+00	6.78E-01	**1.95E+00**	**9.19E-01**	6.14E+00	7.24E-01
f_{12}	**1.40E+01**	**4.21E+01**	4.55E+02	1.30E+03	5.27E+03	6.78E+03	6.15E+02	2.14E+03	1.86E+02	4.79E+02
f_{13}	**1.46E-01**	**1.05E-01**	1.89E+00	7.36E-01	5.81E+00	4.07E+00	1.29E+00	4.13E-01	3.94E+00	9.29E-01
f_{14}	**3.13E+00**	**3.88E-01**	3.20E+00	2.29E-01	3.62E+00	2.22E-01	3.48E+00	2.27E-01	3.39E+00	2.85E-01
f_{15}	**7.96E-07**	**3.26E-07**	2.56E+02	7.60E+01	3.36E+02	9.30E+01	2.03E+02	6.61E+01	2.53E+02	7.43E+01
f_{16}	1.46E+02	2.25E+01	1.72E+02	1.17E+01	1.62E+02	2.83E+01	**1.28E+02**	**1.32E+01**	1.83E+02	1.34E+01
f_{17}	1.49E+02	1.90E+01	1.94E+02	3.51E+01	1.90E+02	2.92E+01	**1.42E+02**	**1.64E+01**	1.95E+02	3.91E+01
f_{18}	4.77E+02	1.45E+02	7.10E+02	1.69E+02	**4.34E+02**	**2.00E+02**	5.80E+02	1.42E+02	6.15E+02	1.04E+02
f_{19}	**4.27E+02**	**1.69E+02**	6.87E+02	1.69E+02	4.91E+02	2.06E+02	5.82E+02	1.41E+02	6.11E+02	1.07E+02
f_{20}	**4.17E+02**	**1.33E+02**	6.77E+02	1.54E+02	4.24E+02	1.95E+02	5.82E+02	1.41E+02	6.11E+02	1.07E+02
f_{21}	**5.75E+02**	**3.13E+01**	7.48E+02	1.28E+02	8.71E+02	3.02E+02	8.08E+02	3.51E+02	7.76E+02	1.86E+02
f_{22}	7.80E+02	3.91E+01	7.79E+02	6.83E+01	7.82E+02	6.99E+00	**7.32E+02**	**1.31E+02**	7.84E+02	4.45E+01
f_{23}	**6.14E+02**	**1.51E+02**	7.81E+02	1.92E+02	6.40E+02	6.68E+01	7.08E+02	2.16E+02	7.12E+02	1.73E+02
f_{24}	2.89E+02	2.32E+02	2.30E+02	1.40E+02	2.43E+02	7.59E+01	**2.00E+02**	**1.64E-02**	2.02E+02	4.06E-01
f_{25}	4.12E+02	1.67E+00	3.99E+02	2.55E+01	4.11E+02	8.09E-01	**3.95E+02**	**4.34E+01**	4.08E+02	9.29E+00

TABLE II. SUCCESS RATE (SR%) AND SUCCESS PERFORMANCE (SP) ACHIEVED FOR PROBLEMS $f_1 - f_{25}$ (d = 10) BY BAS (SP NOT CALCULATED WHEN SR = 0% AND [-] SIGN WAS PUT INSTEAD)

Problem	PLBA		PLIA-BA		Standard BA		Shrinking-based BA		Basic BA	
	SR%	SP	SR%	SP	SR%	SP	SR%	SP	SR%	SP
f_1	**100**	**8.2294E+03**	0	-	0	-	0	-	0	-
f_2	**100**	**4.8090E+04**	0	-	0	-	0	-	0	-
f_3	0	-	0	-	0	-	0	-	0	-
f_4	**100**	**7.1668E+04**	0	-	0	-	0	-	0	-
f_5	**100**	**7.7768E+04**	0	-	0	-	0	-	0	-
$f_6 - f_8$	0	-	0	-	0	-	0	-	0	-
f_9	**100**	**1.8948E+04**	0	-	0	-	0	-	0	-
f_{10}, f_{11}	0	-	0	-	0	-	0	-	0	-
f_{12}	**8**	**1.0497E+06**	0	-	0	-	0	-	0	-
f_{13}, f_{14}	0	-	0	-	0	-	0	-	0	-
f_{15}	**100**	**4.1000E+04**	0	-	0	-	0	-	0	-
$f_{16} - f_{25}$	0	-	0	-	0	-	0	-	0	-

TABLE III. MEAN ERROR VALUES ACHIEVED FOR PROBLEMS $f_1 - f_{25}$ (d = 30) BY BAS

Problem	PLBA		PLIA-BA		Standard BA		Shrinking-based BA		Basic BA	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
f_1	**8.41E-14**	**2.90E-14**	4.96E-02	5.02E-03	2.48E-10	5.02E-11	5.44E-11	6.01E-12	5.00E-02	5.19E-03
f_2	**1.87E-08**	**7.56E-08**	1.81E-01	5.01E-02	1.98E-07	6.76E-07	4.39E+01	2.18E+02	1.81E-01	3.62E-02
f_3	**2.94E+05**	**7.68E+04**	8.16E+05	3.49E+05	2.38E+06	8.06E+05	1.23E+06	4.08E+05	7.80E+05	3.19E+05
f_4	2.87E+03	1.53E+03	5.09E+02	3.32E+02	2.54E+04	8.55E+03	**2.57E+02**	**2.29E+02**	1.50E+04	6.89E+03
f_5	7.24E+03	1.70E+03	4.24E+03	8.03E+02	8.76E+03	2.35E+03	1.10E+04	1.92E+03	**3.93E+03**	**9.76E+02**
f_6	**7.59E+01**	**8.26E+01**	4.40E+02	3.22E+02	6.84E+02	1.34E+03	2.72E+03	4.31E+03	3.36E+02	4.36E+02
f_7	**1.14E-02**	**1.03E-02**	1.12E+00	1.73E-02	8.63E-01	1.88E-01	9.38E-01	3.35E-02	1.12E+00	1.46E-02
f_8	2.01E+01	2.97E-02	**2.00E+01**	**1.92E-02**	2.10E+01	5.85E-02	2.01E+01	4.45E-02	2.01E+01	2.53E-02
f_9	**9.34E-10**	**9.99E-10**	1.93E+02	2.74E+01	1.54E+02	3.51E+01	2.20E+02	3.57E+01	2.25E+02	3.98E+01
f_{10}	**1.29E+02**	**3.08E+01**	2.47E+02	1.81E+01	1.69E+02	7.08E+01	1.56E+02	4.20E+01	2.37E+02	1.73E+01
f_{11}	2.52E+01	3.21E+00	2.65E+01	2.96E+00	3.90E+01	1.40E+00	**1.29E+01**	**3.49E+00**	3.16E+01	1.64E+00
f_{12}	**1.69E+03**	**1.16E+03**	1.73E+04	1.56E+04	6.60E+04	1.63E+05	2.00E+04	1.80E+04	1.16E+04	1.10E+04
f_{13}	**5.59E-01**	**3.02E-01**	1.60E+01	1.92E+00	6.35E+02	2.94E+03	5.33E+00	1.46E+00	2.10E+01	2.19E+00
f_{14}	**1.25E+01**	**3.35E-01**	1.27E+01	2.30E-01	1.30E+01	2.61E-01	1.31E+01	3.80E-01	1.31E+01	2.06E-01
f_{15}	**1.29E+02**	**1.64E+02**	5.76E+02	6.58E+01	4.09E+02	2.77E+01	4.20E+02	3.90E+01	5.78E+02	5.71E+01

f_{16}	1.80E+02	4.44E+01	2.77E+02	3.24E+01	**1.26E+02**	**3.28E+01**	1.91E+02	6.96E+01	2.85E+02	4.10E+01
f_{17}	2.15E+02	8.80E+01	4.60E+02	7.44E+01	**1.80E+02**	**5.43E+01**	2.41E+02	7.30E+01	4.68E+02	7.17E+01
f_{18}	9.11E+02	5.21E+00	9.15E+02	5.63E+00	**9.10E+02**	**2.92E+00**	9.12E+02	1.60E+00	9.16E+02	1.36E+00
f_{19}	**9.09E+02**	2.36E+00	9.16E+02	1.36E+00	**9.09E+02**	**2.32E+00**	9.12E+02	1.99E+00	9.16E+02	1.73E+00
f_{20}	**9.09E+02**	2.50E+00	9.16E+02	1.36E+00	**9.09E+02**	**2.22E+00**	9.12E+02	1.99E+00	9.16E+02	1.73E+00
f_{21}	1.02E+03	2.03E+02	7.18E+02	2.05E+01	**5.91E+02**	**2.02E+02**	7.66E+02	3.14E+02	7.32E+02	1.38E+01
f_{22}	**9.09E+02**	**2.82E+01**	9.44E+02	2.38E+01	9.48E+02	1.43E+01	9.86E+02	5.61E+01	9.41E+02	1.68E+01
f_{23}	1.11E+03	3.05E+00	7.46E+02	2.58E+01	7.03E+02	2.88E+01	**5.57E+02**	**1.25E+01**	7.38E+02	1.71E+01
f_{24}	9.36E+02	1.53E+02	1.36E+03	2.16E+01	2.03E+02	1.57E+01	**2.00E+02**	**0.00E+00**	1.35E+03	2.52E+01
f_{25}	2.13E+02	2.46E+00	2.12E+02	4.93E-01	2.12E+02	8.48E-01	**2.11E+02**	**3.96E-01**	2.13E+02	5.42E-01

TABLE IV. Success Rate (SR%) and Success Performance (SP) Achieved for Problems $f_1 - f_{25}$ (d = 30) by Bas (SP Not Calculated when SR = 0% and [-] Sign was Put Instead)

Problem	PLBA		PLIA-BA		Standard BA		Shrinking-based BA		Basic BA	
	SR%	SP	SR%	SP	SR%	SP	SR%	SP	SR%	SP
f_1	**100**	**7.7945E+04**	0	-	100	2.0265E+05	100	1.8600E+05	0	-
f_2	**96**	2.5518E+05	0	-	96	**2.5216E+05**	64	3.3130E+05	0	-
$f_3 - f_6$	0	-	0	-	0	-	0	-	0	-
f_7	**68**	**1.9660E+05**	0	-	0	-	0	-	0	-
f_8	0	-	0	-	0	-	0	-	0	-
f_9	**100**	**1.1871E+05**	0	-	0	-	0	-	0	-
$f_{10} - f_{14}$	0	-	0	-	0	-	0	-	0	-
f_{15}	**52**	**1.3942E+05**	0	-	0	-	0	-	0	-
$f_{16} - f_{25}$	0	-	0	-	0	-	0	-	0	-

TABLE V. Average Rankings of BA Variants (Friedman) on 10-Dimenstional Problems $f_1 - f_{25}$

Algorithm	Ranking
PLBA	1.66
Shrinking-based BA	2.34
PLIA-BA	3.52
Basic BA	3.64
Standard BA	3.84
p-value	0

TABLE VI. Average Rankings of BA Variants (Friedman) on 30-Dimenstional Problems $f_1 - f_{25}$

Algorithm	Ranking
PLBA	1.76
Standard BA	2.94
Shrinking-based BA	3.04
PLIA-BA	3.42
Basic BA	3.84
p-value	0.000071

TABLE VII. Adjusted p-Values Associated with BA Variants (Friedman) on 10-Dimenstional Problems $f_1 - f_{25}$

Algorithm	Unadjusted *p*	*p Holm*	*p Hochberg*
Standard BA	0.000001	0.000004	0.000004
Basic BA	0.00001	0.000029	0.000029
PLIA-BA	0.000032	0.000064	0.000064
Shrinking-based BA	0.128379	0.128379	0.128379

TABLE VIII. Adjusted p-Values Associated with BA Variants (Friedman) on 30-Dimenstional Problems $f_1 - f_{25}$

Algorithm	Unadjusted *p*	*p Holm*	*p Hochberg*
Basic BA	0.000003	0.000013	0.000013
PLIA-BA	0.000206	0.000617	0.000617
Shrinking-based BA	0.004208	0.008415	0.008326
Standard BA	0.008326	0.008415	0.008326

TABLE IX. SUCCESS RATE (SR%) AND SUCCESS PERFORMANCE (SP) ACHIEVED FOR PROBLEMS $f_1 - f_{25}$ (d = 10) BY PLBA AND OTHER STATE-OF-THE-ART ALGORITHMS (SP NOT CALCULATED WHEN SR = 0% AND [-] SIGN WAS PUT INSTEAD)

Problem	PLBA		DE [46]		SaDE [47]		DMS-PSO [44]		SPC-PNX [45]		Restart CMA-ES [43]	
	SR%	SP	SR%	SP	SR%	SP	SR%	SP	SR%	SP	SR%	SP
f_1	**100**	8.2294E+03	**100**	2.9410E+04	**100**	1.0126E+04	**100**	1.1912E+04	**100**	6.7252E+03	**100**	**1.6100E+03**
f_2	**100**	4.8090E+04	**100**	4.6309E+04	**100**	1.0237E+04	**100**	1.2052E+04	**100**	3.1012E+04	**100**	**2.3800E+03**
f_3	0	-	80	1.1502E+05	64	5.2306E+04	**100**	1.2480E+04	0	-	**100**	**6.5000E+03**
f_4	**100**	7.1668E+04	**100**	5.2372E+04	96	4.5601E+04	0	-	**100**	3.0714E+04	**100**	**2.9000E+03**
f_5	**100**	7.7768E+04	**100**	4.0746E+04	0	-	80	1.1336E+05	**100**	4.0259E+04	**100**	**5.8500E+03**
f_6	0	-	96	4.7398E+04	**100**	4.8777E+04	**100**	5.4677E+04	0	-	**100**	**1.0800E+04**
f_7	0	-	8	1.2000E+06	24	1.7197E+05	16	5.8672E+05	4	1.8033E+06	**100**	**4.6700E+03**
f_8	0	-	0	-	0	-	0	-	0	-	0	-
f_9	**100**	1.8948E+04	44	1.7681E+05	**100**	**1.7048E+04**	**100**	3.4612E+04	0	-	76	7.5700E+04
f_{10}	0	-	0	-	0	-	0	-	0	-	**92**	**6.5000E+04**
f_{11}	0	-	**48**	**1.8852E+05**	0	-	0	-	4	1.0943E+06	24	2.6300E+05
f_{12}	8	1.0497E+06	76	7.1904E+04	**100**	**3.1933E+04**	76	5.4443E+04	0	-	88	3.2700E+04
f_{13}, f_{14}	0	-	0	-	0	-	0	-	0	-	0	-
f_{15}	**100**	4.1000E+04	4	2.4600E+06	92	**3.3165E+04**	88	5.6563E+04	0	-	0	-
$f_{16} - f_{25}$	0	-	0	-	0	-	0	-	0	-	0	-

TABLE X. SUCCESS RATE (SR%) AND SUCCESS PERFORMANCE (SP) ACHIEVED FOR PROBLEMS $f_1 - f_{25}$ (d = 30) BY PLBA AND OTHER STATE-OF-THE-ART ALGORITHMS (SP NOT CALCULATED WHEN SR = 0% AND [-] SIGN WAS PUT INSTEAD)

Problem	PLBA		DE [46]		SaDE [47]		DMS-PSO [44]		SPC-PNX [45]		Restart CMA-ES [43]	
	SR%	SP	SR%	SP	SR%	SP	SR%	SP	SR%	SP	SR%	SP
f_1	**100**	7.7945E+04	**100**	1.3855E+05	**100**	2.0234E+04	**100**	5.0263E+03	**100**	3.0326E+04	**100**	**4.5000E+03**
f_2	96	2.5518E+05	0	-	96	1.4883E+05	**100**	1.2552E+05	88	3.1536E+05	**100**	**1.3000E+04**
f_3	0	-	0	-	0	-	84	3.4100E+05	0	-	**100**	**4.2700E+04**
f_4	0	-	0	-	52	5.3816E+05	0	-	**76**	3.6334E+05	40	**5.9000E+04**
f_5	0	-	0	-	0	-	0	-	0	-	**100**	**6.5900E+04**
f_6	0	-	0	-	0	-	98	3.2781E+05	4	5.2053E+06	**100**	**6.0000E+04**
f_7	68	1.9660E+05	88	1.9952E+05	80	1.3477E+05	96	5.9577E+04	64	3.7063E+05	**100**	**6.1100E+03**
f_8	0	-	0	-	0	-	0	-	0	-	0	-
f_9	**100**	1.1871E+05	0	-	**100**	**9.8934E+04**	0	-	0	-	36	7.9000E+05
f_{10}	0	-	0	-	0	-	0	-	0	-	**12**	**2.4200E+06**
f_{11}	0	-	0	-	0	-	0	-	0	-	**4**	**4.9800E+06**
f_{12}	0	-	0	-	0	-	16	1.5108E+06	0	-	**32**	**2.2500E+05**
f_{13}, f_{14}	0	-		-	0	-	0	-	0	-	0	-
f_{15}	**52**	**1.3942E+05**		-	0	-	0	-	0	-	0	-
$f_{16} - f_{25}$	0	-		-	0	-	0	-	0	-	0	-

TABLE XI. AVERAGE RANKINGS OF PLBA AND OTHER ALGORITHMS (FRIEDMAN) ON 10-DIMENSTIONAL PROBLEMS $f_1 - f_{25}$

Algorithm	Ranking
Restart CMA-ES	3.4
MABC	3.56
DMS-PSO	4.28
SaDE	4.28
DE	4.6
PLBA	**5.02**
ABC	5.12
SPC-PNX	5.74
p-value	0.010674

TABLE XII. AVERAGE RANKINGS OF PLBA AND OTHER ALGORITHMS (FRIEDMAN) ON 30-DIMENSTIONAL PROBLEMS $f_1 - f_{15}$

Algorithm	Ranking
Restart CMA-ES	2.8
SaDE	2.9667
DMS-PSO	3.1667
PLBA	**3.8333**
DE	4.9
SPC-PNX	5.1333
MABC	5.2
p-value	0.001348

TABLE XIII. ADJUSTED p-VALUES ASSOCIATED WITH PLBA AND OTHER ALGORITHMS (FRIEDMAN) ON 10-DIMENSTIONAL PROBLEMS $f_1 - f_{25}$

Algorithm	Unadjusted *p*	*p Holm*	*p Hochberg*
SPC-PNX	0.000731	0.00512	0.00512
ABC	0.013043	0.078255	0.078255
PLBA	**0.019373**	**0.096867**	**0.096867**
DE	0.083265	0.333058	0.333058
SaDE	0.204024	0.612072	0.408048
DMS-PSO	0.204024	0.612072	0.408048
MABC	0.817361	0.817361	0.817361

TABLE XIV. ADJUSTED p-VALUES ASSOCIATED WITH PLBA AND OTHER ALGORITHMS (FRIEDMAN) ON 30-DIMENSIONAL PROBLEMS $f_1 - f_{15}$

Algorithm	Unadjusted *p*	*p Holm*	*p Hochberg*
MABC	0.002346	0.014075	0.014075
SPC-PNX	0.003096	0.01548	0.01548
DE	0.007762	0.031049	0.031049
PLBA	**0.1902**	**0.570599**	**0.570599**
DMS-PSO	0.64205	0.832662	0.832662
SaDE	0.832662	0.832662	0.832662

```
For CurSite = 1 : m
    Set CurBestSite = Sites[CurSite];
    While (not all RecruitBee recruited)
        Set Time = t;
        While(Time != 0)
            Distribute the current recruit bee from the current best site to search the
            neighborhood area of the current site according to Levy flight distribution (3) with search size or scale $\gamma_2$ .
            If(NewSite is better than CurBestSite)
                CurBestSite = NewSite ;
                Break;
            End If
            Time = Time – 1;
        End While
    End While
    Sites[CurSite] = CurBestSite;
End For
Shrink the Levy search size ( $\gamma_2$ ) by a shrinking factor ( $sf$ )
```

Fig. 1. Pseudo-code of the proposed Greedy Levy-based Local Search Algorithm (GLLSA)

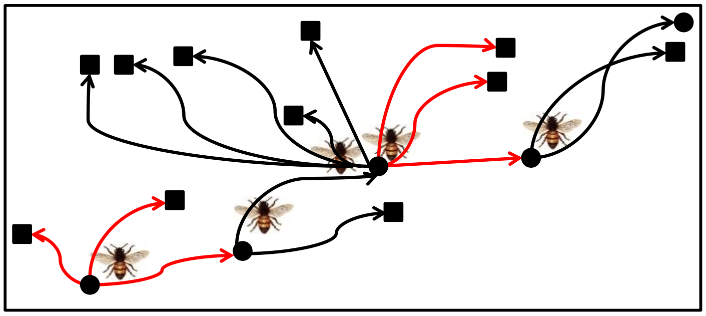


Fig. 2. A schematic diagram of recruit bees foraging and exploiting a patch in the proposed local search algorithm (● for better site than the current best site and ■ for worse sites)

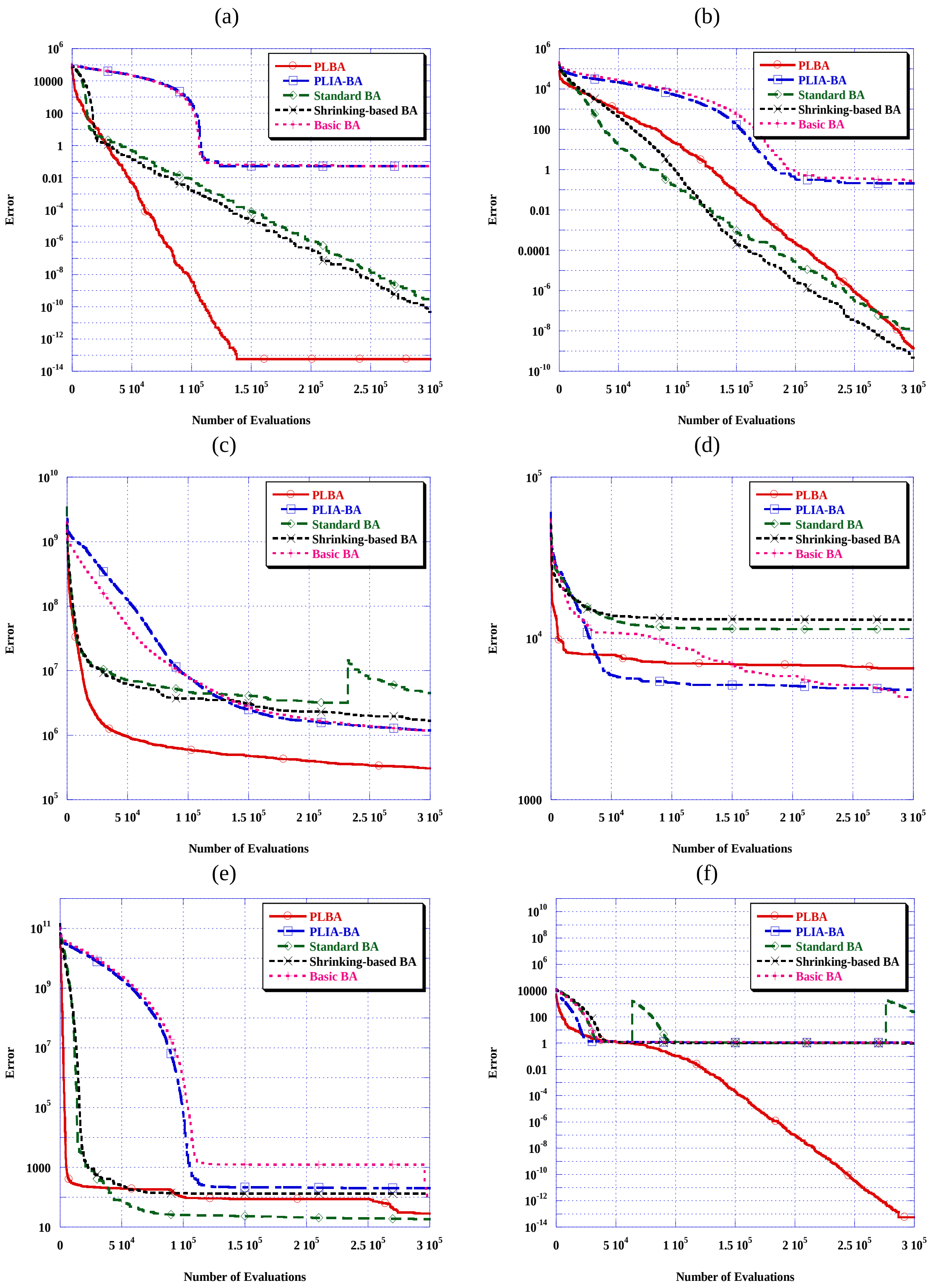


Fig. 3. Convergence behavior of the BA variants on functions: (a) f_1, (b) f_2, (c) f_3, (d) f_5, (e) f_6, and (f) f_7

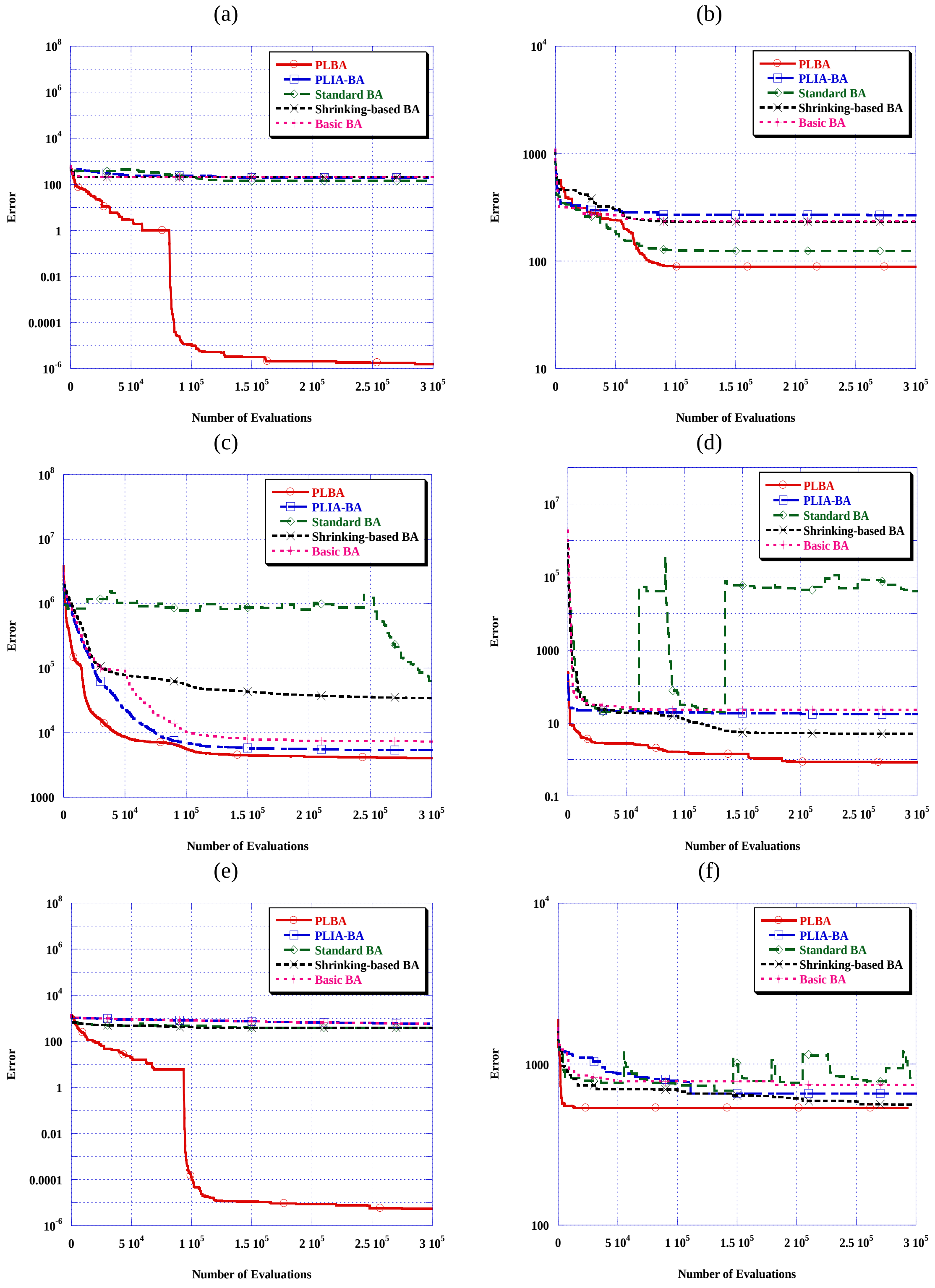


Fig. 4. Convergence behavior of the BA variants on functions: (a) f_9, (b) f_{10}, (c) f_{12}, (d) f_{13}, (e) f_{15}, and (f) f_{23}

APPENDIX A

Table A. 1 The parameters used with different values for different problems in Basic BA, Shrinking-based BA, and Standard BA

Function	Basic BA	Shrinking-based BA		Standard BA		
	ngh	ngh_{init}	*sf*	ngh_{init}	*sf*	*stlim*
f_1	0.1	1	0.999	1	0.999	700
f_2	0.1	1	0.999	1	0.999	700
f_3	0.1	1	0.9999	1	0.800	700
f_4	1	5	0.999	5	0.999	700
f_5	1	1	0.9991	1	0.9991	700
f_6	0.1	1	0.999	1	0.999	700
f_7	5	5	0.9999	5	0.9999	700
f_8	1E-3	0.001	0.999	1E-3	0.999	700
f_9	0.1	0.1	0.999	1	0.999	700
f_{10}	1	1	0.999	1	0.999	700
f_{11}	0.1	0.1	0.999	1	0.999	700
f_{12}	0.01	0.01	0.9999	1	0.999	700
f_{13}	0.1	1	0.999	1	0.999	700
f_{14}	0.1	0.1	0.999	1	0.999	700
f_{15}	1E-3	3	0.999	3	0.999	700
f_{16}	1	1	0.999	3	0.999	700
f_{17}	0.01	1	0.999	3	0.999	700
f_{18}	1	1	0.99991	3	0.999	700
f_{19}	1	1	0.99991	3	0.999	700
f_{20}	1	1	0.99991	3	0.999	700
f_{21}	1	1	0.990	1	0.990	700
f_{22}	1	1	0.999	3	0.999	700
f_{23}	1	1	0.9999	3	0.9999	700
f_{24}	0.1	1	0.999	3	0.999	700
f_{25}	1	1	0.9999	3	0.9999	700

Table A. 2 The parameters used with different values for different problems in PLIA-BA, and PLBA

Function	PLIA-BA			PLBA					
	ngh	*P*	γ_1	*P*	γ_1	γ_{2_init}	γ_3	*t*	*sf*
f_1	0.1	1	1E-7	19	3	1E-2	1E-2	5	0.985
f_2	0.1	1	1E-7	1	1	1	1	10	0.990
f_3	0.1	1	1E-7	1	1	1E-3	1	10	1
f_4	1	1	1E-7	5	3	2	1	60	0.960
f_5	1	1	1E-7	1	1	2	1	50	0.950
f_6	0.1	1	1E-7	1	1	1	1	20	0.990
f_7	5	20	1E-7	10	1	4	1	30	0.990
f_8	1E-4	1	1E-7	1	1	1E-5	1	50	1
f_9	1E-3	1	1E-7	10	4	7E-2	1E-7	45	0.960
f_{10}	1	1	1E-7	1	1	1	1	50	0.980
f_{11}	0.01	19	1	1	1	1	1	20	0.980
f_{12}	0.01	1	1E-7	17	1	1E-4	3	60	1
f_{13}	0.1	19	1E-7	12	1E-3	1E-4	1E-3	40	1
f_{14}	5	1	1E-7	1	1	1E-2	1	50	1
f_{15}	1E-3	1	1E-7	1	1E-7	4E-5	1E-2	30	1
f_{16}	1	1	1E-7	1	1	1	1	30	0.990
f_{17}	0.01	1	1E-7	1	1	1	1	50	0.990
f_{18}	1	1	1	19	1	2	1	20	1
f_{19}	1	1	3	19	3	2	3	15	0.990
f_{20}	1	1	3	19	1	1	1	20	0.990
f_{21}	1	1	1	19	3	1	3	15	1
f_{22}	1	19	1	1	1	1	1	15	0.990
f_{23}	1	1	1	1	1	1	1	20	0.999
f_{24}	0.1	1	1	1	5	1	5	40	0.980
f_{25}	1	2	1	1	5	2	5	40	0.980

Comprehensive Centralized-Data Warehouse for Managing Malaria Cases

Nova Eka Diana
YARSI E-Health Research Center
Faculty of Information Technology
YARSI University
Jakarta, Indonesia

Aan Kardiana
YARSI E-Health Research Center
Faculty of Information Technology
YARSI University
Jakarta, Indonesia

***Abstract*—Tanah Bumbu is one of the most endemic areas in Indonesia for patients diagnosed with malaria diseases. Currently, available malaria case data were stored in disparate sources. Hence, it is difficult for the public health department to quickly and easily gather the useful information for determining strategic actions in tackling these cases. The purpose of this research is to build a data warehouse that integrates all malaria cases from disparate sources. This malaria data warehouse is a centralized architecture of galaxy or constellation scheme that consists of three fact tables and 13 dimension tables. SQL Server Integration Services (SSIS) is utilized to build ETL packages that load data from various sources to stages, dimensions, and fact tables in malaria data warehouse. Finally, a timely report can be generated by extracting the salient information located in malaria data warehouse.**

Keywords—malaria case; centralized data warehouse; galaxy scheme; ETL; timely report

I. Introduction

Malaria is one of infectious disease spread by a mosquito of the genus Anopheles. This animal carries out a plasmodium parasite and spread it into human blood circulation through a bite. Every year, there are about 300-500 million people infected by malaria and 1 million people died because of this disease. According to Global Health Observatory (GHO) data, in 2012, there were an estimated about 207 million cases of malaria worldwide and most of those (about 80%) occurred in sub-Saharan Africa. In this report, Indonesia, one of the endemic countries, was mentioned to have about 343,527 cases of malaria with 45 people reported to be dead in 2013 [1]. The number of malaria incidences in Indonesia has decreased by 2.9% in 2007 to 1.9% in 2013. However, contrary to these conditions, the number of cases in West Papua has increased quite sharply in 2013. This area is located in the east part of Indonesia with the most prevalence number above the average [2]. One possible reason for these phenomena is the fairness of medical supply distribution. The scattering locations of malaria incidence may affect a different responsiveness over the cases treatments. Lack of centralized malaria data all over Indonesia could be the reason as well.

Data warehouse has been widely used to manage a significant volume of data that spread in scattered locations. Data warehousing system can be considered as a collection of methods, techniques, and tools to assist managerial users, e.g. senior managers and directors, to administer their jobs. Data warehouse could provide some salient information that helps these users to conduct data analysis in decision-making processes [3]. In recent years, healthcare industry and organization have started adopting a predictive analytic approach for a variety of purposes. To support this idea, they must develop an infrastructure that able to generate timely reports and intervention strategies for health care problems. Healthcare organization needs to build an advanced data warehouse that integrates all available information in a real time manner so that can accommodate those capabilities. A proper deployment of successful data warehouse in managing diseases information will benefit both the organizations and the patients [4-5].

In this article, data warehouse building is proposed to manage and integrate a scattered data of Malaria cases in Tanah Bumbu, one of the endemic areas in Indonesia. By using this proposed data warehouse, the public health department can extract and generate information that useful for decision-making processes. They also can generate and visualize timely reports to present information in an interactive way that easier for executives in the public health department to understand.

The rest of this article is organized as following. Section 2 describes literature review about methodologies used to develop a data warehouse. Next, Section 3 explains the centralized architecture and Single Dimensional Data Store (DDS) scheme used to build malaria data warehouse. In Section 4, data collection and analysis are conducted to determine the stages, dimensions, and fact tables for malaria data warehouse. After that, ETL packages for translating various data sources into a single data warehouse are depicted in Section 5. Finally, the conclusion and future works are defined in Section 6.

II. Data Warehouse Methodology

Data warehouse can be considered as a central repository of information that integrates various data from one or more disparate sources. Data warehouse is usually described as a collection of subject-oriented, integrated, non-volatile, and time-varying data to support decision-making processes [6]. Data warehouse gives a multidimensional view of a big amount of historical data from operational data sources to provide useful information for decision maker in improving their organization business process [7].

This research is funded by Unggulan Perguruan Tinggi (UPT) grant from Directorate General of Higher Education (DIKTI) of Indonesia Ministry of Education and Culture.

WH Inmon made an observation on classical system development life cycle (SDLC) which assumes that requirements are identified at the beginning of design process of Decision Support System (DSS). However, in the real practices, requirements are usually the last one discovered in development process [6]. Inmon pointed out that data warehouse design was a data-driven approach of which analyst frequently understood requirements and data that available for them, only after they encountered opportunities to perform various kinds of data analysis.

Data warehouse development process is a cycle rather than a serialized time and it repeats every 12 to 18 months [8]. This cycle consists of five major steps as illustrated in Fig. 1.

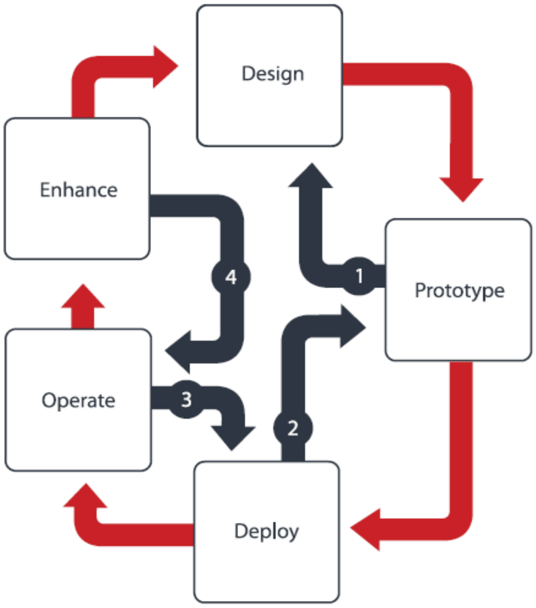


Fig. 1. Data Warehouse Lifecycle Model [8]

Those five major steps are:

- Design. In this phase, the developers create a robust dimensional data models based on available data and analyst requirements.
- Prototype. The main objective of this stage is to constrain and reframe end-user requirements by giving a group of decision-makers and leading practitioners what they need.
- Deploy. There are at least two separate deployment processes should be conducted in the development process: deployment of a prototype to production-test environment and performance-tested production to an actual production environment.
- Operate. Here, the developer conducted a day-to-day maintenance of the data warehouse.
- Enhance. This step includes modification of physical components, operations and management processes in response to business requirement changes.

III. Data Warehouse Infrastructure

A. Data Warehouse (DW) Architecture

Five dominant architectures that are usually adopted to build DW infrastructure: independent data mart, hub-and-spoke, bus, centralized and federated architecture [9-11]. Among these architectures; bus, hub-and-spoke, and centralized, are equally successful for their intended purposes. There is no single dominant architecture in terms of information and system quality, individual and organizational impacts. There is no clear winner among these architecture designs. The differences among them lie on cost, adaptability, scalability, and efficiency of an organizational business process. Therefore, IT managers should consider those factors in deciding the right architecture for building their data warehouse infrastructure [12-13].

In this research, the centralized architecture of malaria data warehouse was proposed as illustrated in Fig. 2. Many applications can directly access this data warehouse to extract salient information for advanced purposes. The whole processes executed in centralized architecture are depicted in Fig. 3. Data warehouse system collects raw data from various data sources such as a database (DBMS), spreadsheet and CSV files. All of these data are then transformed into a uniform format and stored in one place called as Data Stage. After that, these data will be distributed to various components of data warehouse storage, e.g. dimensional and fact tables. When disparate data sources have resided in malaria data warehouse, then different applications can connect to provide any services related to decision support making.

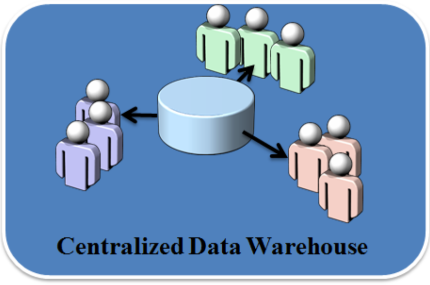


Fig. 2. Centralized Data Warehouse

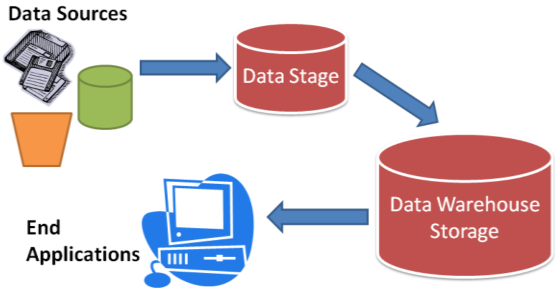


Fig. 3. Processes in centralized architecture

B. Data Flow Architecture

Here, Single Dimensional Data Store (Single DDS) was employed to build malaria data warehouse as shown in Fig. 4.

In this data flow, ETL packages gathered data from various source systems and move it to Stage data store. Next, DDS ETL packages and Data Quality (DQ) would extract data from Stage data store and distribute it to a correct DDS in the data warehouse. Data resided in DDS were then accessed to provide useful information for various applications. Throughout these processes, control-audit administered all ETL packages based on the structure of the data and the description of the processes saved in a metadata.

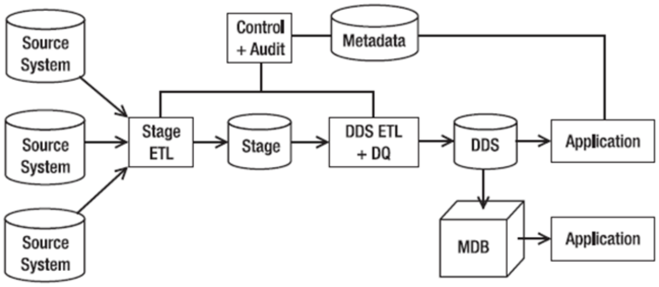


Fig. 4. Single DDS Data Flow Architecture [14]

IV. Data Collection and Analysis

Data sources used in this research were a collection of patients' data diagnosed with malaria in Tanah Bumbu Regency, South Kalimantan, one of the most endemic areas in Indonesia. These data sources consisted of numerous spreadsheet files that recorded malaria cases from 2012 to 2014. Each file contained information about the status of patients' medication for one month period. Therefore, there would be twelve spreadsheet files that recorded malaria cases in a year. Moreover, the medication process for a patient diagnosed with malaria might need more than one month. Hence, it was also possible for one patient' data to be recorded in more than one spreadsheet file.

Here, malaria data were analyzed to identify entities, attributes, and relationship among the entities. Based on the data behavior, those entities were categorized into dimension and fact tables. In this process, numerous spreadsheet files were transformed into a uniform format and then located in a staging table. There were two stage tables created in this data warehouse, Stage of Malaria Data and Stage of Area Information. Stage of Malaria Data recorded patients' data that were diagnosed with malaria. Stage of Area Information stored data about areas and location of the health services that managed Malaria cases.

Fig. 5 illustrates the relationship among stages, dimensions, and fact tables of the malaria data warehouse. Here, more than one fact table and one-dimensional table are shared by many fact tables. Hence, it is called as Galaxy or Fact-Constellation Scheme. In this malaria data warehouse, there are three fact tables: Fact of Malaria Data, Fact of Area Statistic, and Fact of Logistic. Fact of Malaria Data records rows of data related to patient information, type of malaria disease, laboratory results, and type of services given to patients. Fact of Area Statistic contains data that can be used to analyze the starting point and the spread of malaria disease. Fact of Logistic records medicines and supplies that are distributed to each health services that responsible for handling patients diagnosed with malaria. This table can also be used to analyze and predict the utilization of some medicines and supplies for each health services location in each period. Three fact tables, Fact of Malaria Data, Fact of Area Statistic, and Fact of Logistic tables, shared Geographic Dimensional table to track the location of patient and health services area.

V. Results and Discussion

After designing the scheme structure of Malaria data warehouse, data sources from numerous spreadsheet files were converted and formatted into a uniform format. Multiple ETL packages were created using SQL Server Integration Services (SSIS) to populate data to malaria data warehouse. These ETL packages were classified into three categories: Stage ETL, Dimension ETL, and Fact ETL packages.

A. Stage ETL Packages

Stage ETL packages were created to populate data sources into stage tables in the data warehouse. Stage tables are merely just a regular tables, but they have a role in storing rows of data from various data sources. Hence, they can be accessed quickly. Two ETL packages were created for populating data into StageDataMalaria and StageInfoWilayah from spreadsheet data sources. Fig. 6 and Fig. 7 depicted data flow process of Stage of Malaria Data and Stage of Area Information ETL. Stage of Malaria Data ETL consisted of three components: data source, data flow transformation, and data destination. Data flow transformation in this ETL was a derived column type to generate a new value by applying some expressions. ISNULL expression was utilized here to convert number into string values as shown in Table I.

TABLE I. Data Conversion Example

Value	Generated Value
NULL	""
1	PUSTU
2	Poskesdes
3	Polindes/Bidan Desa
4	Klinik/Praktek Swasta
5	Kader/Posmaldes

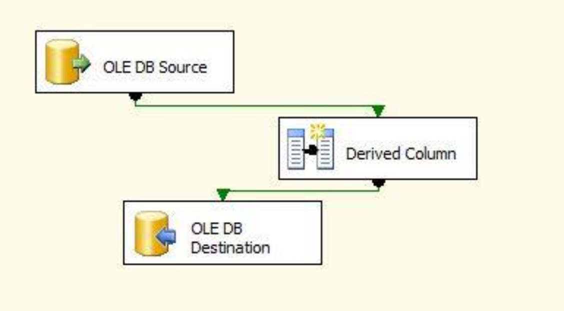


Fig. 5. Stage of Malaria Data ETL

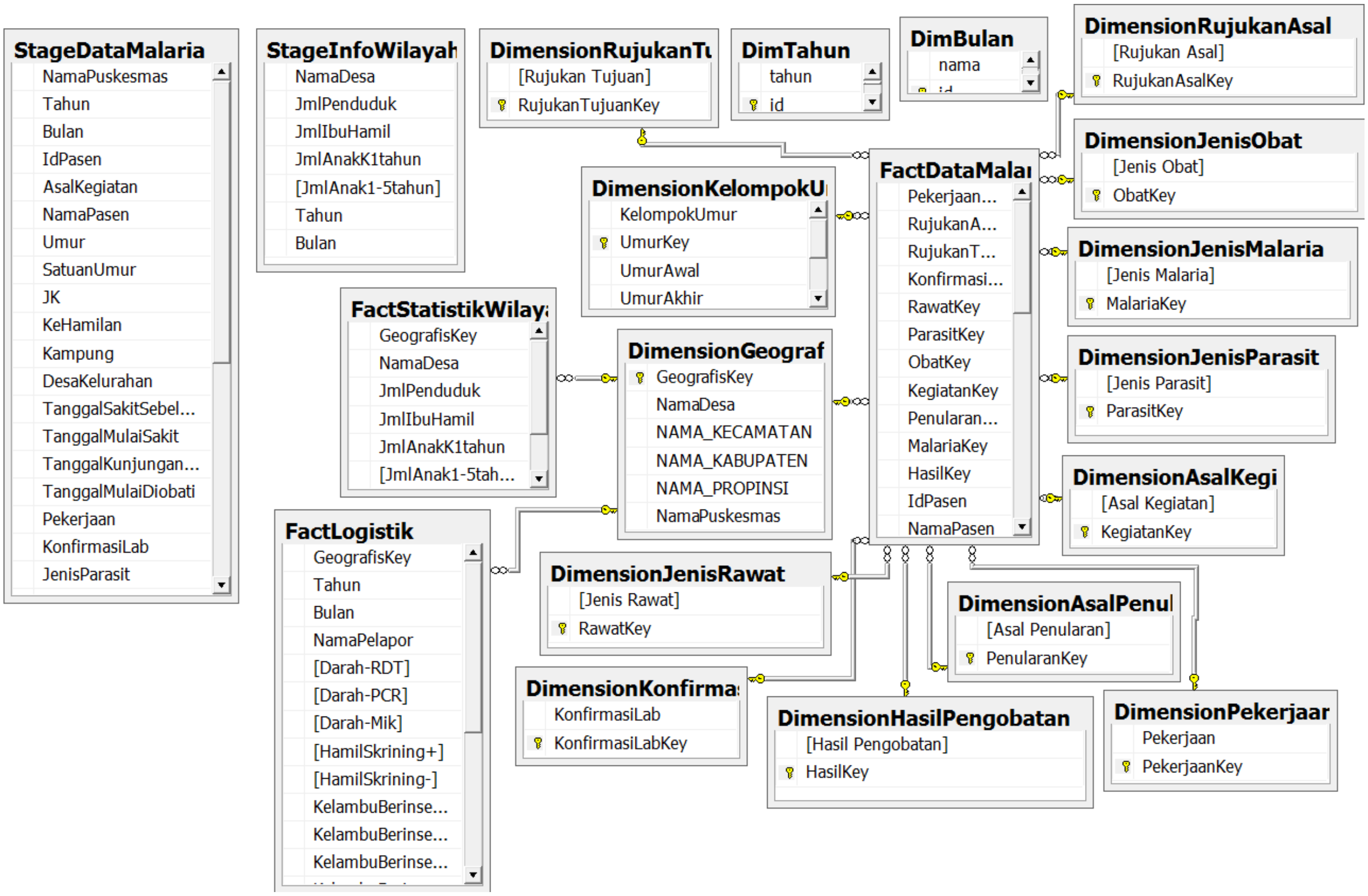


Fig. 6. Galaxy Scheme of Malaria Data Warehouse

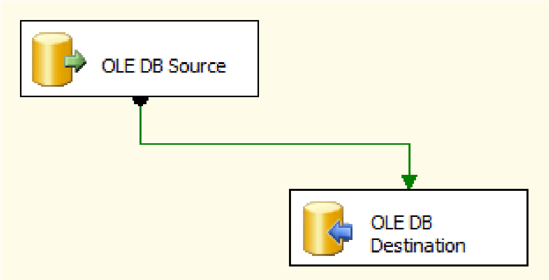


Fig. 7. Stage of Area Information ETL

B. Dimensional ETL Packages

All dimensional ETL packages transform the master data sources into the dimensional tables. Data from dimensional and stage tables will be utilized by ETL packages to fill in the fact tables in the data warehouse. In this process, there are 13 ETL packages created to fill in dimensional tables in Malaria data warehouse as depicted in Table II.

Fig. 8 illustrates the process of executing ETL to populate data into Dimension of Medication Result. In this process, ETL connects to master data sources, extract the medication result and then populate them into Dimension of Medication Result. Fig. 9 shows data in Dimension of Medication Result after successfully execute the ETL. ETL package is successfully executed when the color of each component is green and displays the number of data transferred from OLE DB Source to OLE DB Destination. In this case, four rows of data are successfully populated into Dimension of Medication Result.

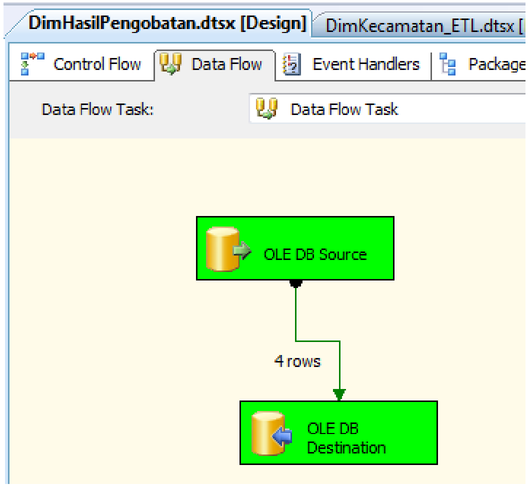
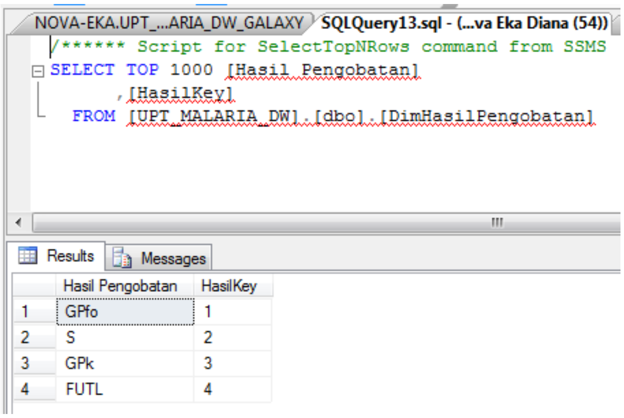


Fig. 8. Dimension of Medication Result ETL execution

TABLE II. DIMENSIONAL ETL PACKAGES

No	ETL Name
1	Dimension of Destination Reference ETL
2	Dimension of Age Group ETL
3	Dimension of Geography ETL
4	Dimension of Treatment Type ETL
5	Dimension of Lab Confirmation ETL

No	ETL Name
6	Dimension of Medication Result ETL
7	Dimension of Spread Starting Point ETL
8	Dimension of Occupation ETL
9	Dimension of Activity Starting Point ETL
10	Dimension of Parasite Type ETL
11	Dimension of Malaria Type ETL
12	Dimension of Medicine Type ETL
13	Dimension of Starting Point Reference ETL


NOVA-EKA.UPT_...ARIA_DW_GALAXY
SQLQuery13.sql - (...va Eka Diana (54))
Results
Messages

```
/****** Script for SelectTopNRows command from SSMS
SELECT TOP 1000 [Hasil Pengobatan]
      ,[HasilKey]
  FROM [UPT_MALARIA_DW].[dbo].[DimHasilPengobatan]
```

	Hasil Pengobatan	HasilKey
1	GPfo	1
2	S	2
3	GPk	3
4	FUTL	4

Fig. 9. Result of DimHasilPengobatan ETL execution

C. Fact ETL packages

ETL packages in this group are used to fill in the fact tables by extracting data from dimension and stage tables. Three ETL packages are created to populate data into Fact of Malaria Data, Fact of Area Statistic, and Fact of Logistic. Each of these ETL consists of two components, OLE DB Source and OLE DB Destination. To fill in these fact tables, OLE DB Source is created as JOIN SQL command between dimensional and stage tables. For example, Fig. 10 illustrates SQL command that is used to build OLE DB Source and also the preview result to fill in Fact of Malaria Data table.

D. Malaria Case Reports

Malaria data warehouse can be utilized to generate important information for high-level management in the Public Health Department. Users at this level can make a quick and precise decision to overcome Malaria spread by accessing information pooled in the data warehouse.

Fig. 11 depicts a preview of reporting application that use data warehouse information. This figure illustrates the number of Malaria cases based on the type of Malaria Parasite. On January 2013, number of patients in Batulicin1 area diagnosed with Plasmodium Falciparum, Plasmodium Vivax, and Mixed parasite, were 10%, 51%, and 36%, respectively. Based on this information, the health department can decide about what kind of treatments and medicines should be given to those areas. Fig. 12 also gives a preview of reporting application that can be generated from malaria data warehouse by giving some parameters. Here, patients are classified based on their primary occupation. Occupation information is important to find the possible reason of Malaria spread. If most of the patients diagnosed with Malaria have an occupation as farm workers, then the possible reason is the after harvest condition of farming areas that have not been recovered yet. Otherwise, if most of the patients are a miner, then perhaps the responsibility parties have not taken any action to restore the field to its previous condition.

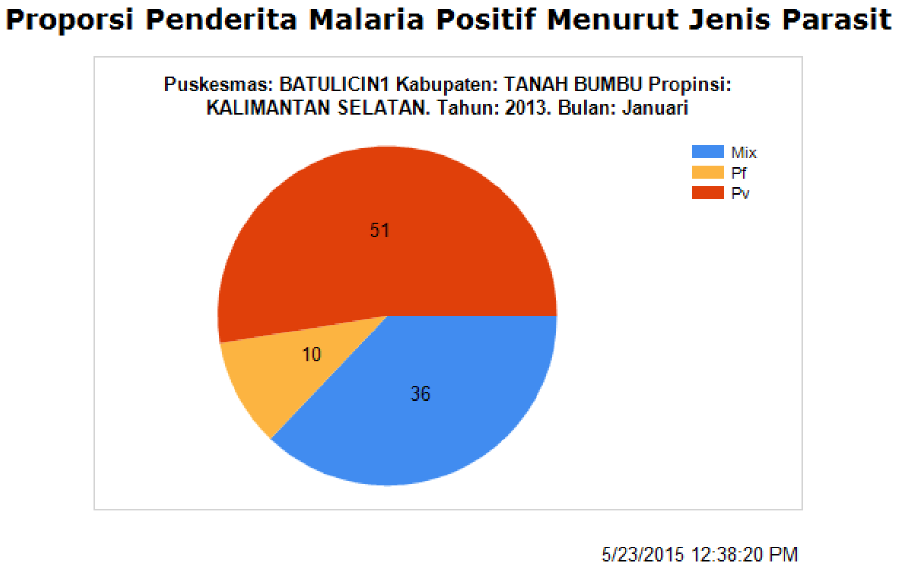
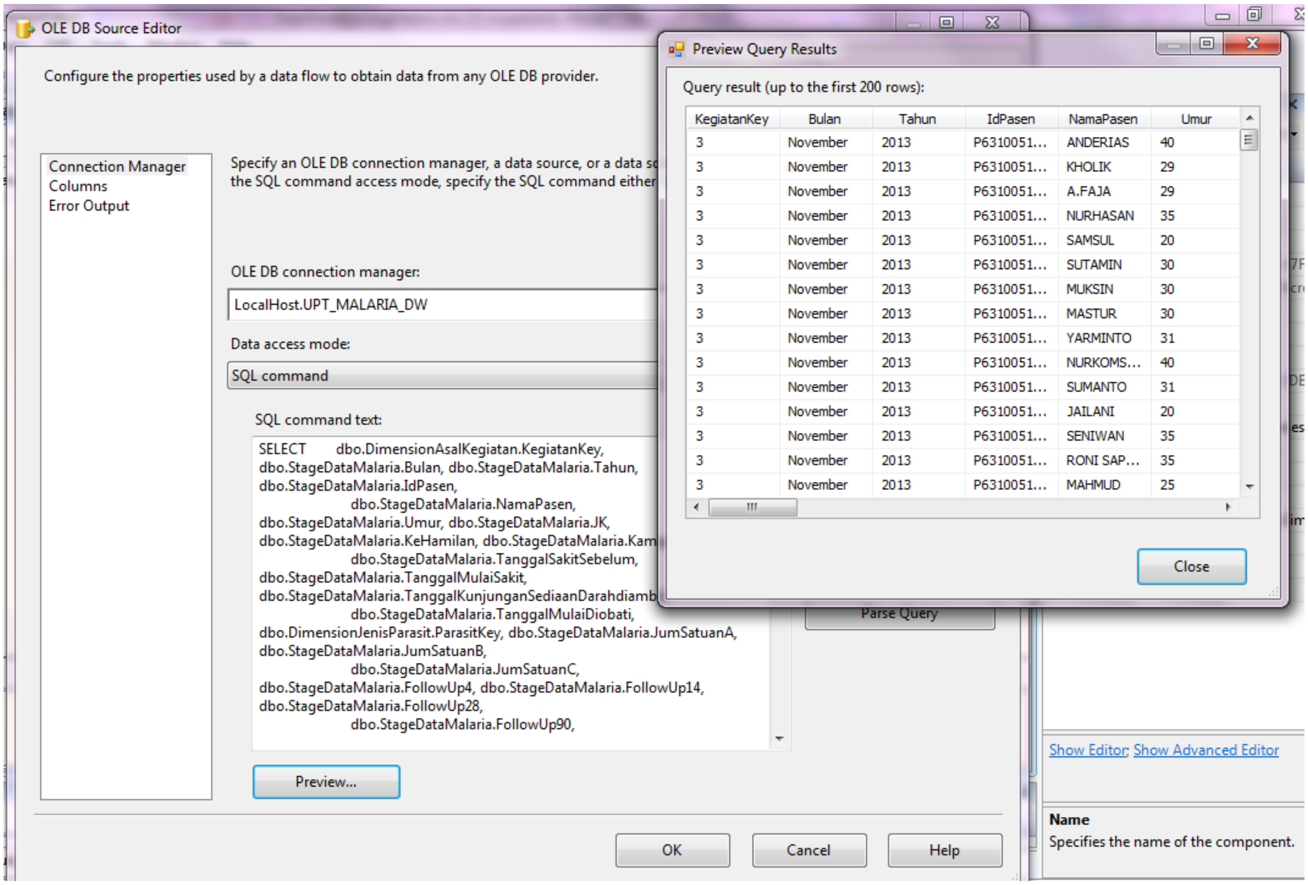


Fig. 10. Result of DimHasilPengobatan ETL execution



OLE DB Source Editor

Configure the properties used by a data flow to obtain data from any OLE DB provider.

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data so
the SQL command access mode, specify the SQL command either

OLE DB connection manager: LocalHost.UPT_MALARIA_DW

Data access mode: SQL command

SQL command text:

```
SELECT      dbo.DimensionAsalKegiatan.KegiatanKey,
dbo.StageDataMalaria.Bulan, dbo.StageDataMalaria.Tahun,
dbo.StageDataMalaria.IdPasen,
                        dbo.StageDataMalaria.NamaPasen,
dbo.StageDataMalaria.Umur, dbo.StageDataMalaria.JK,
dbo.StageDataMalaria.KeHamilan, dbo.StageDataMalaria.Kam
                        dbo.StageDataMalaria.TanggalSakitSebelum,
dbo.StageDataMalaria.TanggalMulaiSakit,
dbo.StageDataMalaria.TanggalKunjunganSediaanDarahdiamb
                        dbo.StageDataMalaria.TanggalMulaiDiobati,
dbo.DimensionJenisParasit.ParasitKey, dbo.StageDataMalaria.JumSatuanA,
dbo.StageDataMalaria.JumSatuanB,
                        dbo.StageDataMalaria.JumSatuanC,
dbo.StageDataMalaria.FollowUp4, dbo.StageDataMalaria.FollowUp14,
dbo.StageDataMalaria.FollowUp28,
                        dbo.StageDataMalaria.FollowUp90,
```

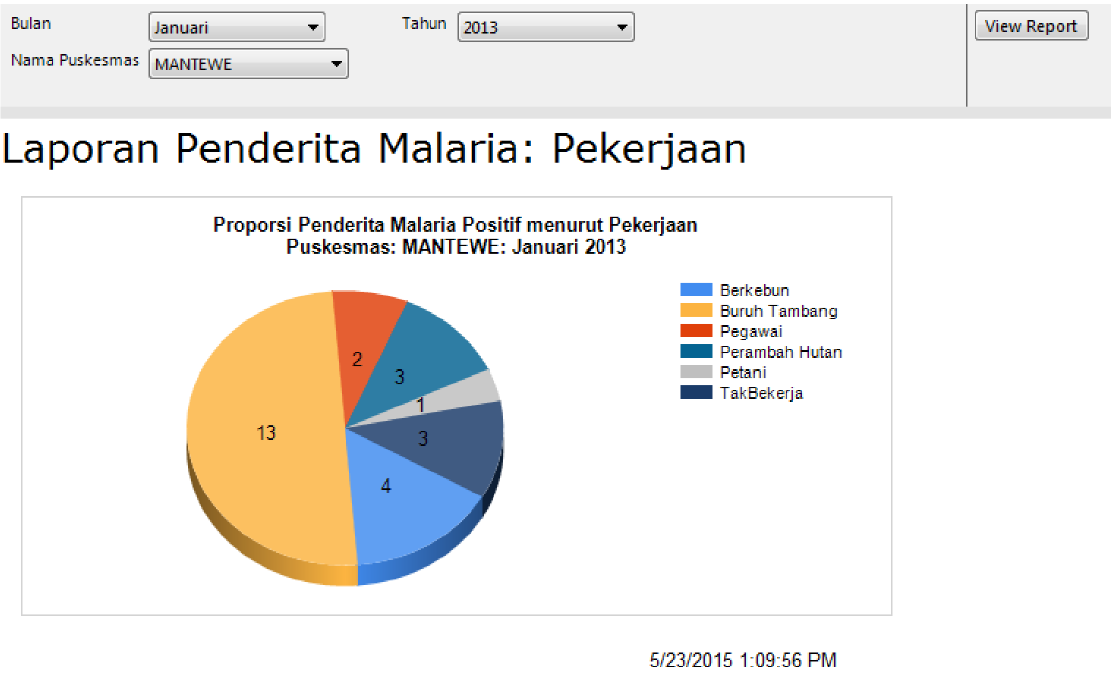
Preview Query Results

Query result (up to the first 200 rows):

KegiatanKey	Bulan	Tahun	IdPasen	NamaPasen	Umur
3	November	2013	P6310051...	ANDERIAS	40
3	November	2013	P6310051...	KHOLIK	29
3	November	2013	P6310051...	A.FAJA	29
3	November	2013	P6310051...	NURHASAN	35
3	November	2013	P6310051...	SAMSUL	20
3	November	2013	P6310051...	SUTAMIN	30
3	November	2013	P6310051...	MUKSIN	30
3	November	2013	P6310051...	MASTUR	30
3	November	2013	P6310051...	YARMINTO	31
3	November	2013	P6310051...	NURKOMS...	40
3	November	2013	P6310051...	SUMANTO	31
3	November	2013	P6310051...	JAILANI	20
3	November	2013	P6310051...	SENIWAN	35
3	November	2013	P6310051...	RONI SAP...	35
3	November	2013	P6310051...	MAHMUD	25

Name
Specifies the name of the component.

Fig. 11. OLE DB source & result Preview for Fact DataMalaria



Bulan: Januari
Tahun: 2013
Nama Puskesmas: MANTEWE

Laporan Penderita Malaria: Pekerjaan

Proporsi Penderita Malaria Positif menurut Pekerjaan
Puskesmas: MANTEWE: Januari 2013

5/23/2015 1:09:56 PM

Fig. 12. Reports of Malaria patient’s occupation

VI. CONCLUSION

In this study, a comprehensive data warehouse is proposed for managing and extracting the salient information about Malaria cases in Tanah Bumbu. This malaria data warehouse is centralized architecture of galaxy or constellation scheme that consists of 3 fact tables and 13 dimension tables. SSIS engine is employed to build ETL packages that load data from various data sources into a stage, dimension, and fact tables in the malaria data warehouse. Timely reports, i.e. number of patients diagnosed with malaria report, can be generated by extracting information from Fact of Malaria Data and Geographic Dimension.

This malaria data warehouse is a foundation to integrate all malaria cases from various endemic and non-endemic areas in Indonesia. Future work would create and generate data mining rules to know the possible starting point where malaria case happened. Hence, the related official government may take an appropriate action to overcome this problem. Furthermore, information existed in the data warehouse is also important for biologists to determine the habitat and the season where mosquito of the genus Anopheles lives. Therefore, the public health could make some preventive actions for minimizing an upcoming number of malaria cases.

REFERENCES

[1] WHO, “World malaria report 2013,” WHO Global Malaria Programme, World Health Organization, 2013.

[2] Riskesdas, “Riset kesehatan dasar,” Badan Penelitian Dan Pengembangan Kesehatan, Kementrian Kesehatan RI, 2013.

[3] Golfarelli M. & Rizzi S., Data warehouse design: modern principles and methodologies, McGraw-Hill, 2009.

[4] Alazmi A. R., “Data warehousing implementations: a review,” International Journal on Data Mining and Intelligent Information Technology Applications (IJMIA), Vol. 4, No. 1, 2014.

[5] Ramick D.C., “Data warehousing in disease management programs”, J. Healthc Inf Manag, Vol. 15, pp. 15:99-105, 2001.

[6] Inmon W. Building the data warehouse, Wiley Technology Publishing, Wiley, 2005.

[7] Pardillo C.J. and Mazon J.N., “Using ontologies for the design of data warehouses,” Int. J. Database Manag. Syst., vol. 3, no. 2, pp. 73–87, May 2011.

[8] Demarest M., “Understanding the data warehouse lifecycle,” WhereScape White Paper, 2013.

[9] Nilakanta S., Scheibe K. and Rai A., “Dimensional issues in agricultural data warehouse designs,” Computer and Electronics in Agriculture, Vol. 60, No. 2, pp. 263-278, 2008.

[10] Sen A. and Sinha A.P., “A comparison of data warehousing methodologies using a common set of attributes to determine which methodology to use in a particular data warehousing project,” Communications of the ACM, Vol. 48, No. 3, pp. 79-84, 2005.

[11] Singh S. and Malhotra S., “Data warehouse and its methods,” Journal of Global Research in Computer Science, Vol. 2, No. 5, pp.113-115, 2011.

[12] Ariyachandra T. and Watson H.J., “Which data warehouse architecture is most successful?,” Business Intelligence Journal, Vol. 11, No. 1, pp. 4-6, 2006.

[13] Alazmi A. R., “Data warehousing implementations: a review,” International Journal on Data Mining and Intelligent Information Technology Applications (IJMIA), Vol. 4, No. 1, 2014.

[14] Rainardi V., Building a data warehouse with examples in sql server,” Apress, pp. 29-48, 2008.

An Adaptive Approach to Mitigate Ddos Attacks in Cloud

Baldev Singh
Research Scholar
IKG Punjab Technical University
Jalandhar, India

S.N. Panda
Director (Research)
Chitkara University
Rajpura, India

***Abstract*—Distributed denial of service (DDOS) attack constitutes one of the prominent cyber threats and among the hardest security problems in modern cyber world. This research work focuses on reviewing DDOS detection techniques and developing a numeric stable theoretical framework used for detecting various DDOS attacks in cloud. Main sections in the paper are devoted to review and analysis of algorithms used for detection of DDOS attacks. The framework theorized here deals with the variability calculation method in conjunction with sampling, searching methods to find a current state of a particular parameter under observation for detecting DDOS attacks. This way a solution is to build that measure the performance and conduct the monitoring framework to capture adversity related to DDOS attacks. The described algorithm intends to capture the current context value of the parameters that determine the reliability of the detection algorithm and the online pass algorithm helps to maintain the variability of those collected values thus maintaining numerical stability by doing robust statistical operations at endpoints of traffic in cloud based network.**

***Keywords*—DDOS attack; Intrusion detection; Threshold; Cloud; virtual machine**

I. Introduction

Internet of Things (IoT) has evolved in modern times in leap and bounds. Incidents of attacks over the Internet especially DDOS attacks [1] are increasing day by day. News are dominated by successful DDOS attacks against a number of major Fortune 100 organizations in this arena. Hence, many security experts have being thinking on how does Organized Fraud Rings (OFR's) make money out by disrupting the information technology (IT) assets of a company? How can we detect and disrupt this activity? [28,29] Historically, the fragmented organizations that unwittingly constitute this "fraud supply chain" could provide no coherent indication of suspicious activity. Each step could only be observed in isolation. However, solutions are now available to correlate events across the multiple disparate systems involved to join the dots and see the complete picture.

Centralized collection of activities through every phase of the disruptive activity like DDOS attack [1] process can identify and alert on suspicious patterns and stop it in its tracks, even where evasive techniques such as geographical dispersion and Internet service provider (ISP) switching are employed. The purpose of the DDOS attack [3] is normally about breaking the business continuity. Hence, there is a need to build a mechanism on identifying the correlation of activity and transaction velocity of DDOS Attack. The existing network security mechanisms confront new challenges in the cloud [2] such as virtual machine intrusion attacks [4,5] and malicious user activities. Hence, new security methods [6] are required to increase users' level of trust in clouds. Presently, cloud service providers implement data encryption for the data centers, virtual firewalls and access control lists.

The number of threat vectors in today's cloud landscape [7] is increasing day by day that is the problem. The existing detection approaches are not foolproof. There is need to learn more about attackers and adopt an adaptive approach to defense at every stage. This can happen only if the algorithm of detections must be adaptive and numerically stable. There is need to focus on methods on how to gather threat Intelligence about our adversaries and know the motives of malicious attacker?

The cloud based networks [16] and services are prone to suffer from malicious attacks because of their inherent characteristics of being accessible globally any time and also due to the frequent changes in topology and development of IoT as well as because of landscape nature of Internet. Of particular concern, it is the denial of service attacks that makes the service unavailable to its intended cloud users. In fact, there are three major techniques which are: misuse detection, anomaly detection and specific detection like DDOS attack. Each of these detection methods has its pros and cons. These are however, reviewed in the Related Work section. But from the current incidence reports it is clear that new combinational methods need to be implemented for the proper working of cloud industry. It is known fact that both the internal and external anatomy [8] of the data-center matters, how it is structured architecturally to measure the volume of traffic is also the main critical point, if somehow the intruders are able to launch a slow attack it must be detectable or if it is a sudden flood of packets then the system must be able to mitigate the flood to have clean traffic. This is not possible unless there is continuous monitoring which includes the mapping of threats [18] cope with the understanding correlations of all the factors contributing to the adversity. Therefore, the thresholds of finding inflection points where the traffic changes to malicious is essential to successful running of data centers in cloud in thwarting the DDOS attacks.

The rest of this paper is as follows. Section II studies a couple of related works in detection and defense mechanism of DDOs attacks. In Section III, the main research gaps in the reviewed work are highlighted. Section IV contains the threshold calculation mechanism required for DDOS attack detection. Section V summarizes the shortcomings of already contributed work. In Section VI, some concluding remarks are given and finally the future scope of study is briefed in Section VII.

II. Related Work

Due to mammoth size of network and IoT, there are various types of vulnerabilities in cloud based network. The impact of adversity on the cloud based network is not easy to estimate due the dynamic attributes of the system dynamics. These dynamics force us to think on developing various mathematical models based on which the adversity may be captured. Since, all these methods are based on mathematical model of DDOS threat detection and when they are put to test against the real life scenarios, their performance comes into question. Hence, this area of research explores many possibilities to mitigate the DDOS attacks. The Intrusion detection systems (IDS) [24, 25, 37] were first implemented as frugal, optional mitigation mechanisms but with Big Data technologies the direction of the industry to implement a permanent robust solutions at all ends of the network including ISP and customer end is now propagated. This leads to scenario where normal detection techniques using static values rendered becomes useless and as per our systematic review of related works in the field of intrusion/malicious attack detection [9,18], there are variety of mechanisms that can be used to detect anomalies or malicious behavior in the cloud based network. Each of these techniques/mechanisms discussed here are threshold basis and have their own pros and cons. A summary of some related literature that holds trade-offs in its favour, is presented in this section.

In [10] Static threshold based intrusion detection systems were proposed by Kim et al. (2004), Gates and Damon (2005), Leckie et al (2002) and Faizal et al (2009). Network IDS technique proposed by (Abdollah, Masud, Shahrin & Robiah, 2009) has used the concept of static threshold, although dynamic threshold is better solution. Threshold value is selected to take decision in identifying the DDOS attacks. It is the basic unit that differentiates between the normality and abnormality in the traffic over the cloud based network. In this static threshold method, a cutoff is fixed which decides the normal and abnormal levels. The basic features used in detection of attacks are mainly IP address of the victim machine, timestamp, time-duration of connection, protocol used, connection status flag, and source & destination services.

The threshold is used to differentiate between normal traffic and abnormal traffic [17] in the network. This threshold value is acquired by using observation and experimental technique and the verified by using statistical process control approach [13]. Although static threshold methods are easy to implement for detection as well and have low computational complexity but the main weakness of this method is that fix value range is never close to real systems and are changing with respect to time. This method does not take into account the differential or cumulative threshold which given better response to adversity in real time.

In [15] Proactive DDOS attack detection and defense mechanisms propound by Keromytis, Misra & Rubenstein 2002. It focuses on the advance detection of attacks. In reactive mechanisms, detection the attacks is by using signatures (attack pattern) or anomaly behavior. Proactive mechanisms emphasize on to improve the reliability of the global Internet infrastructure by adding additional functionality to Internet components to detect and defend various attacks as well as vulnerability exploitation. The main objective of this approach is to make the infrastructure resistant to the attacks and to provide service to normal users continually under extreme conditions.

Reactive defense approach was suggested by Ioannidis & Bellovin 2002. Third-party Intrusion Detection Systems (IDS) are deployed to obtain attack related information and then action is taken according to this information. Various strategies are used for intrusion detection purpose. If the IDS system unable to detect the DDoS attack packets, then filtering mechanisms are used, that are able to filter out the attack stream completely, even at the source network. A rate limit is imposed by the IDS on the stream to characterize that the stream is malicious.

In [13], Dynamic threshold based approach is proposed by Gupta, Sanchika, Padam Kumar, and Ajith Abraham. [13] Paper addresses to vulnerabilities responsible for known attacks on cloud and analyzed various measures to secure cloud based network from these malicious attacks both insider and outsider. A profile is created for each machine to detect and prevent various cloud attacks. In this approach, the estimation of upper and lower cutoff points are figured after some time period, after every time slot, the value may change. However, there is not much difficulty in implementation of various methods of calculations like method based on average, mode, frequency, deviation from mean. But this method cannot handle extreme high and low values statistically that may cause to wrong calculations (numerically unstable) as change point detection method may calculate wrong threshold, thereby increase the false alarm rate.

[11] B. B. Gupta, R. C. Joshi, M. Misra proposed "Dynamic and Auto Responsive Solution for Distributed Denial-of-Service Attacks Detection in ISP Network," In this paper, they presented a framework which emphasize on characterization of a wide range of flooding DDOS attacks [32] and their detection. As per [11] flooding attacks are because of high rate disruptive, diluted low rate degrading and varied rate and there should be monitoring of the propagation of abrupt traffic changes inside ISP network. [11] suggested that a profile of the traffic normally seen in the network is to be created, and then anomalies are detected whenever traffic goes out of profile. Although they claimed that the said detection system is scalable to varying network conditions and adapts itself to different attacks [19] loads but to identify threshold values detection mechanism is not free from generating false positive alarm rates and hence not foolproof for detection of malicious flows characterization.

In [37], C. Modia, D. Patela, B. Borisaniyaa, H. Patelb, A. Patelc, and M. Rajarajanc in their paper, "A survey of intrusion detection techniques in cloud," are really of the opinion that the existing Intrusion Detection Systems and Intrusion Prevention Systems for cloud environments are short of feasible solution. The authors explained that the explored solutions of IDS are still far from the integration in the clouds. They suggested that these must be combined with security information and event management in addition to adopting additional security measures and correlation rules to identify internal attacks or to be prepared for zero-day attacks. They made stress on the necessity of the centralized view in monitoring IDS and making advancements in existing solutions with competence to examine data stream by scaling up and down.

[14] Jun-Ho Lee, Min-Woo Park, Jung-Ho Eom, and Tai-Myoung Chung in their paper “Multi-level intrusion detection system and log management in cloud computing” proposed multi-level IDS in combination with log management approach to strengthen the security in cloud based network so that anomaly behavior can be detected in cloud environments. In this approach they are of the view that more rules and patterns are proportionate to the strength of the security in cloud computing provided that proper logs are to be administered. It is based on the quantification of risk levels and assigning risk points in proportion to risk anomaly behavior. Although the approach is fully quantitative in nature which is significant in any corporate security risk program but the main thing is that more the rules and patterns, more will be the complex and slow IDS in cloud computing.

III. Research Outline

The major gaps are found on the basis of review study and its findings are as follows:

- It is difficult to judge the point of inflection or point change which can reflect the abnormal behavior due to DDOS attack.
- It is difficult to sometime identify the end points of cloud from where the DDOS attack is creating problem.
- It is hard to identify the intensity of attack, if the attack is slow in nature and has discrete events occurring based on demand cloud services.
- It is difficult to calculate accurately the thresholds of parameters that can reflect DDOS attack behavior as ranges may changing with patterns difficult to comprehend.

Pseudo code to detect DDOS attacks in Cloud based network: Following are the proposed steps to detect DDOS attacks in cloud based network:

- Measure normal traffic Volume
- Measure normal traffic Flow
- Upper bound of Threshold value for Volume
- Measure threshold value for Flow
- Measure lower bound of threshold value for Volume
- If attack pattern is detected due to flooding attacks, generate DDOS attack alert.

IV. Mechanism to Compute Threshold for Ddos Detection

An ‘abnormality’ may be found while observing a particular network factor. It is either the values of parameters that start touching abnormal ‘lows’ or ‘highs’ at certain ‘intervals’ of data series or values of parameters start scaling higher values from the normal scale. There is variance [36] which indicates the abnormality. It is more suitable to use online algorithm in cases where cost of computation is sensitive to the response time of an operation. Moreover, the method of (SEM) Structured Equation Modeling [23] which is mainly used to test conceptual/theoretical model, intends to identify an equation that defines a sequence of values between two or more parameters under observations for DDOS detection, based on time series sampling and then further sample for threshold value based methods.

(a1) *Pseudo code of Online Algorithm to compute variance for n samples:*

```
For defined timeline;
N=0; Mean=0;
M2=0;
Data=[];
For x in Data
N=N+1;
Delta=x-Mean;
Mean=Mean+ Delta/n;
M2=M2+Delta*(x-Mean)
If(n<2)
Return 0;
Variance=M2/(n-1)
End.
```

(a2) *Empirical rule:* The empirical rule [20,21] (Three Sigma rule) states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

- 68% of data falls within the first standard deviation from the mean.
- 95% fall within two standard deviations.
- 99.7% fall within three standard deviations.

Check ‘Empirical Rules as if:

a) 68% values within the first standard deviation values
b) 95% values within two standard deviation values
c) 97% values within three standard deviation
then attack flag=’true’ else attack flag=’false’;

(a3) *Area Elimination Method*

Let there be 2 points x_1 and x_2 which lie in the interval (a,b) of the sample extracted at any given time for analysis and satisfy ($x_1<x_2$), hence the three rules:

i. If $f(x_1)>f(x_2)$ then the threshold does not lie in (a,x_1)

ii. If $f(x_1)<f(x_2)$ then the threshold does not lie in (x_2,b)

iii. If $f(x_1)=f(x_2)$ then the threshold does not lie (a,x_1) and (x_2,b).

First Rule Scenario:

If the function value at x_1 is larger than that at x_2, threshold point x cannot lie on the left side of x_1.

Second Rule Scenario:

If the function value at x_1 is less than x_2, the threshold point 'x' cannot lie in right side.

Third Rule Scenario:

When $f(x_1)==f(x_2)$, implies that there is one lowest or highest value that can be taken as threshold. Hence R_1 and R_2 areas are eliminated.

Golden Ratio method

Golden Ratio method [20], may be used as golden ratio pattern is also found in nature extensively, the 'abnormal' threshold values may be captured using golden ratio interval division method.

The sample search space (a,b) is first linearly mapped to a unit interval search space (0,1) and then two data points [22] at τ from either end of the search space are chosen so that at every iteration, the elimination region ($1-\tau$) to that in the previous iteration is covered. This can be achieved by equaling ($1-\tau$) with (τ x τ). This yields the golden number $\tau=0.618$.

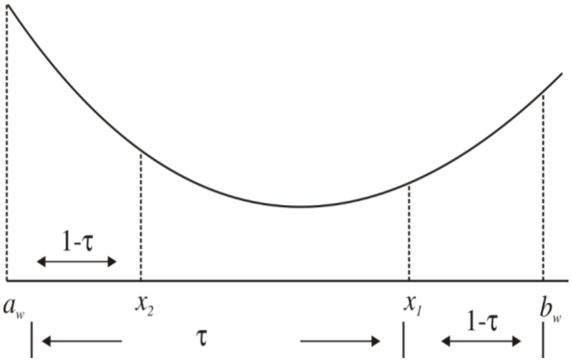


Fig. 1. Golden Ratio points (x1 and x2)

Algorithm: DDOS data under observations is divided into golden ratio parts.

Step 1:

Choose a lower bound *a* and an upper bound *b*. using Synthetic Division method [21,22]. Set small value ϵ. Normalize the variables 'x' by using the equation

$\omega=(x-a)/(b-a)$.

Thus $a\omega=0$, $b\omega=1$, and $\pounds\,\omega=1$, set k=1.

Step 2:

Set $\omega 1=a\omega+(0.618)\ \pounds\omega$

$\omega 2=b\omega-(0.618)\ \pounds\ \omega$.

Compute $f(\omega 1)$ or $f(\omega 2)$

Apply the basic Area elimination method [a3].

Step3 :

Is $|\pounds\omega|<\epsilon$

If no, set k=k+1, goto Step 2;

Else Terminate.

Step 4:

Online algorithm as in section [a1]

Step 5:

Check 'Empirical Rules' as in section [a2]

If large no of outlier form

Attack='true'

Else

Attack='false'

End;

The number of function evaluation 'n' required for achieving a desired accuracy, ϵ is calculated by solving the following equation

$$(0.618)^{n-1}=(b-a)<\epsilon$$

V. Discussion

There is plethora of methods to overcome the issue of rise in DDOS attacks and have understood that it affects their overall eco-system of conducting 'Business on Internet or Cloud'. These algorithms offer various degrees of stability, reliability in their working. The latest findings, occurrences, incidence reports [33] on DDOS attacks suggest that these methods still need a deeper examination as inadequate interpretations are done. Bandwidth [30] is one way to measure DDOS attacks including application layout attacks. Our solution also checks periodically the DNS TTL of all the DNS associated with our cloud endpoints. The ratio of abnormal data to the normal data will normally be imbalanced so the golden method tries to compensate this imbalance. The proposed method as discussed above is also better as real time information of deviation may not be possible every time due to inherent nature of data stream analyzed at any given time. This method is based on capturing abnormality at certain logical intervals. If a particular 'current' threshold or current value of the objective function, defined by well defined relationship, occurs above 'normal range' successively in multiple intervals during evaluation, there is high possibility that the network is going under 'adversity'. Care has been taken to evaluate the successive function evaluation values whose average does not go far away from the mean using robust online algorithm. The method has advantage in the sense that it checks the well defined mathematical relationship between two variables that influence the performance of network under ‘adversity’ using structure equation modeling.

VI. CONCLUSION

Today we have entered a world where cyber enemy politics [31], cyber terrorism, cyber hacktivism have taken over cyber crime space or attack landscape. Now, DDOS attacks can also be launched from phone, in fact nearly any kind of device with IP address can launch DDOS attacks. Network level DDOS attacks typically require IP-threat level assessment strategy to safeguard against DDOS attacks. Reflective DDOS attacks[12] need ‘state flow’ awareness strategy for Outbound DDOS attacks and in case of Bi-direction flood detection [26], the strategy with algorithms having numerical stability is required. In fact, our proposed algorithm can also cover those kinds of DDOS attacks which involve ‘specially crafted ‘packet attacks, where protocol analysis reveals the correct situation in conjunction with threshold techniques discussed in Section-III**.** Other than the DDOS attacks that involve Recon (Scan) [26] or that involve highly advanced evasion, detection requires further advanced methods in this context. The latest Gartner report suggests that application layer level DDOS attacks account for maximum incidences (25%) which requires behavioral analysis techniques due to multifold increase in DDOS attacks[33]. It is also evident that today’s service provider solutions cannot avoid analysis of protocol behaviour, volume, type of traffic for its survival. In summary, the proposed work covers following aspects to break into the anatomy of DDOS attacks. It covers the numerically stable calculations of DDOS attacks that impact network level Flow Volumetric attacks [38].

1) Proposed method covers the attacks initiated as “randomized” ,‘slow’ and ‘low’ [30,35] at application layer. The method proposed do calculations based on application’s traffic end points also, where, ‘payload’ matching or entropy [27] based methods become useless due to high pattern variations.

2) Deploy multi stage algorithm [34] that can block spectrum of unwanted traffic as well as dynamic unwanted users. This way detection becomes more effective.

3) However, support from other components of defense mechanism shall also be solicited which include keeping an eye on application users and unwanted activities or simply enforcing usage standard, enforcement of protocol anomalies and violation by enforcing RFC and industry standards.

VII. FUTURE SCOPE

The proposed solution can easily be integrated into big data analysis based solutions as it allows scalability where application of machine learning algorithms may also help. For future work, it is suggested that SEM may be used for threat modeling of DDOS attacks along with column based data mining algorithms, which use partial probability theory for detection of DDOS attacks in cloud based network.

REFERENCES

[1] Bing Wang, Yao Zheng, Wenjing Lou, Y. Thomas Hou, “DDoS attack protection in the era of cloud computing and Software-Defined Networking,” Computer Networks, Volume 81, 22 April 2015, pp 308-319, ISSN 1389-1286.

[2] Iankoulova, I., Daneva, M. “Cloud computing security requirements: A systematic review” in Research Challenges in Information Science (RCIS), 2012 Sixth International Conference; 2012.

[3] DDoS attacks on the rise – by criminals and spies, Network Security, Volume 2014, Issue 2, February 2014, Page 2, ISSN 1353-4858.

[4] Danny Bradbury, “The problem with Bitcoin,” Computer Fraud & Security, Volume 2013, Issue 11, November 2013, Pages 5-8.

[5] Steve Mansfield-Devine, The evolution of DDoS, Computer Fraud & Security, Volume 2014, Issue 10, October 2014, Pages 15-20, ISSN 1361-3723.

[6] P. Varalakshmi, S. ThamaraiSelvi, “Thwarting DDoS attacks in grid using information divergence”, Future Generation Computer Systems, Volume 29, Issue 1, January 2013.

[7] Pitropakis Nikolaos, Anastasopoulou Dimitra, Pikrakis Aggelos and Lambrinoudakis, Costas,” If you want to know about a hunter, study his prey: detection of network based attacks on KVM based cloud environments”, Journal of Cloud Computing,Vol 3, 2014.

[8] Ben-Porat, U.; Bremler-Barr, A.; Levy, H., "Vulnerability of Network Mechanisms to Sophisticated DDoS Attacks," IEEE Transactions on Computers 2013, , vol.62, Issue 5.

[9] Dit-Yan Yeung, Yuxin Ding,“User Profiling for Intrusion Detection Using Dynamic and Static Behavioral Models,” Springer Berlin Heidelberg, 2002.

[10] Faizal M.A., Zaki M.M., Shahrin S., Robiah Y., and Rahayu S.S., (2010) “Statistical Approach for Validating Static Threshold in Fast Attack Detection,” Journal of Advanced Manufacturing Technology, Vol. 4, 2010.

[11] B. B. Gupta, R. C. Joshi, M. Misra, "Dynamic and Auto Responsive Solution for Distributed Denial-of-Service Attacks Detection in ISP Network," International Journal of Computer Theory and Engineering; pp. 71-80, 2009.

[12] WeiWei; Feng Chen; Yingjie Xia; GuangJin, "A Rank Correlation Based Detection against Distributed Reflection DoS Attacks," Communications Letters, IEEE , vol.17, no.1, pp.173,175, January 2013.

[13] Gupta, Sanchika, Padam Kumar, and Ajith Abraham. "A profile based network intrusion detection and prevention system for securing cloud environment."*International Journal of Distributed Sensor Networks* 2013 (2013).

[14] L. Jun-Ho, p. min-Woo, E. Jung-Ho, C. Tai-Myoung, ” Multi-level intrusion detection system and log management in cloud computing”, *Proceedings of the 13th International Conf. pp. 552–555, Feb.* 2011.

[15] Guangsen Zhang and Manish Parashar , “Cooperative Defence against DDoS Attacks” Journal of Research and Practice in Information Technology, 2006.

[16] S. Qaisar and K. Khawaja, “Cloud computing: network/security threats and countermeasures”, Interdisplinary Journal of Contemporary Research In Business Volume 3, January 2012.

[17] Kollias, S.; Vlachos, V.; Papanikolaou, A.; Chatzimisios, P.; Ilioudis, C.; Metaxiotis, K., "A global-local approach for estimating the Internet's threat level," Journal of Communications and Networks, vol.16, no.4, pp.407,414, Aug. 2014.

[18] Vasanthi, S.; Chandrasekar, S. (2011), “A study on network intrusion detection and prevention system current status and challenging issues”. *Advances in Recent Technologies in Communication and Computing* (ARTCom 2011), 3rd International Conference on , vol., no., pp.181,183, 14-15.

[19] Monowar H. Bhuyan, H. J. Kashyap, D. K. Bhattacharyya and J. K. Kalita (2013),”Detecting Distributed Denial of Service Attacks: Methods, Tools and Future Directions”, *The Computer Journal.*

[20] Deb, Kalyanmoy: Optimization for engineering design: Algorithms and examples”. PHI Learning Pvt. Ltd., 2012.

[21] Aufmann, Richard, and Joanne Lockwood: *Introductory and Intermediate Algebra: An Applied Approach.* Cengage Learning, 2013.

[22] Scherer, Philipp OJ. "Roots and Extremal Points," In Computational Physics Springer International Publishing”, pp. 83-111, 2013.

[23] Hoyle, Rick H.: *Handbook of structural equation modeling*. Guilford Publications, 2014.

[24] Xinya Wu; Yonghong Chen, "Validation of Chaos Hypothesis in NADA and Improved DDoS Detection Algorithm," *Communications Letters, IEEE*, vol.17, no.12, pp.2396-2399, December 2013.

[25] Chun-Jen Chung, Khatkar, P., Tianyi Xing, Jeongkeun Lee, Dijiang Huang, "NICE: Network Intrusion Detection and Countermeasure Selection in Virtual Network Systems," IEEE Transactions on Dependable and Secure Computing, vol.10, no.4, pp.198,211, July-Aug. 2013

[26] Wang Jin, Zhang Min, Yang Xiaolong, Long Keping, Xu Jie, "HTTP-sCAN: Detecting HTTP-flooding attack by modeling multi-features of web browsing behavior from noisy web-logs," China Communications, vol.12, no.2, pp.118,128, Feb. 2015

[27] Xinlei Ma; Yonghong Chen, "DDoS Detection Method Based on Chaos Analysis of Network Traffic Entropy," IEEE Communications Letters, vol.18, no.1, pp.114,117, January 2014

[28] Shui Yu; Yonghong Tian; Song Guo; Wu, D.O., "Can We Beat DDoS Attacks in Clouds?" IEEE Transactions on Parallel and Distributed Systems, vol.25, no.9, pp.2245,2254, Sept. 2014

[29] Anwar, Z.; Malik, A.W., "Can a DDoS Attack Meltdown My Data Center? A Simulation Study and Defense Strategies," IEEE Communications Letters, vol.18, no.7, pp.1175,1178, July 2014

[30] Geva, M.; Herzberg, A.; Gev, Y., "Bandwidth Distributed Denial of Service: Attacks and Defenses," Security & Privacy, IEEE , vol.12, no.1, pp.54,61, Jan.-Feb. 2014

[31] Paulo Shakarian, Jana Shakarian and Andrew Ruef, Chapter 6 - Cyber Attacks by Nonstate Hacking Groups: The Case of Anonymous and Its Affiliates, In Introduction to Cyber-warfare, edited by Paulo Shakarian, Jana Shakarian, Andrew Ruef, Syngress, Boston, 2013, Pages 67-110.

[32] Jingtang Luo, Xiaolong Yang, Jin Wang, Jie Xu, Jian Sun, Keping Long, "On a Mathematical Model for Low-Rate Shrew DDoS," IEEE Transactions on Information Forensics and Security , vol.9, no.7, pp.1069,1083, July 2014

[33] Akamai's Prolexic Security Engineering and Research Team (PLXsert)," Four-fold increase in DDoS attacks," Network Security, Volume 2014, Issue 11, November 2014, Page 2, ISSN 1353-4858, http://dx.doi.org/10.1016/S1353-4858(14)70107-2.

[34] Fei Wang, Hailong Wang, Xiaofeng Wang, Jinshu Su, "A new multistage approach to detect subtle DDoS attacks, Mathematical and Computer Modelling", Volume 55, Issues 1–2, January 2012, Pages 198-213, ISSN 0895-7177, http://dx.doi.org/10.1016/j.mcm.2011.02.025.

[35] C. Balarengadurai, S. Saraswathi, "Comparative Analysis of Detection of DDoS Attacks in IEEE 802.15.4 Low Rate Wireless Personal Area Network", Procedia Engineering, Volume 38, 2012, Pages 3855-3863, ISSN 1877-7058, http://dx.doi.org/10.1016/j.proeng.2012.06.442.

[36] "Algorithms for calculating variance"; https://en.wikipedia.org/wiki/Algorithms_for_calculating_variance

[37] C. Modia, D. Patela, B. Borisaniyaa, H. Patelb, A. Patelc, and M. Rajarajanc, "A survey of intrusion detection techniques in cloud," Journal of Network and Computer Applications, vol. 36, issue 1, pp. 42–57, 2013.

[38] Jisa David, Ciza Thomas, "DDoS Attack Detection Using Fast Entropy Approach on Flow- Based Network Traffic", Procedia Computer Science, Volume 50, 2015, Pages 30-36, ISSN 1877-0509

[39] S.N. Panda, Singh Baldev, "Defending Against DDOS Flooding Attacks- A Data Streaming Approach", International Journal of Computer & IT (Print Journal), 2015.

Common Radio Resource Management Algorithms in Heterogeneous Wireless Networks with KPI Analysis

Saed Tarapiah
Telecommunication Engineering Dept.
An-Najah National University Nablus,
Palestine

Kahtan Aziz
College of Engineering Computing.
Al Ghurair University Dubai, United Arab
Emirates

Shadi Atalla
Lavoro Autonomo
(LA) Torino, Italy
Italy

***Abstract*—The rapid increase of number of personal wireless communication equipped devices boosts the user service demands on wireless networks. Thus, the spectrum resource management in such networks becomes an important topic in the near future. Notwithstanding, typically, users equipped with multiple wireless interfaces, thus the access operational scenario is no longer based on single Radio Access Technology (RAT). In this work, we studied the heterogeneous wireless communication scenarios, as a joint cooperative management of different RATs through which network providers can satisfy as possible as wide variety of user services demands in a more efficient manner by exploiting their varying characteristics and properties. To achieve this objective, a Common Radio Resource Management (CRRM) algorithms and techniques are proposed and designed to efficiently manage and optimize the radio resources in a heterogeneous wireless networks. In this context, this work studies and analyzes some common radio resource management techniques to efficiently distribute traffic among the available radio access technologies while providing adequate quality of service levels under heterogeneous traffic scenarios. The most interesting algorithms have been critically analyzed and then some in depth investigations with attention on implementations and techno-economic issues are performed on some of the identified CRRM algorithms.**

Keywords—heterogeneous wireless networks; Radio Resource Management (RRM); radio access technology (RAT)

I. INTRODUCTION

Recently, it is clear the wide deployment of several coexistence radio access technologies (RATs), such deployment for wireless scenarios introduces a new dimension to improve and utilize the performance and the efficiency when several radio access technologies are deployed together in comparison to the scare available radio resources.

Several RATs reflect the heterogeneity concept, where this scenario is composed of different Radio Access Network (RAN) each RAN interfacing a Common Core Network (CN). RANs can consist in different cellular networks, e.g. Universal Terrestrial Radio Access Network (UTRAN) either Frequency Division Duplexing (FDD) or Time Division Duplexing (TDD), GSM EDGE Radio Access Network (GERAN), as well as other public non-cellular broadband wireless hotspots, e.g. WLAN IEEE 802.11g or IEEE 802.11n [1]. Typically, The infrastructure of core network is divided into the Packet Switch (PS) and Circuit Switch (CS) domains. The CM provides access to external networks such as Public Switched Telephone Network (PSTN) or the Internet. Recently, the mentioned external networks include other public and private Wireless Local Area Networks (WLANs) providing an interface for terminals to access to the core network services. For more details refer to figure 1.

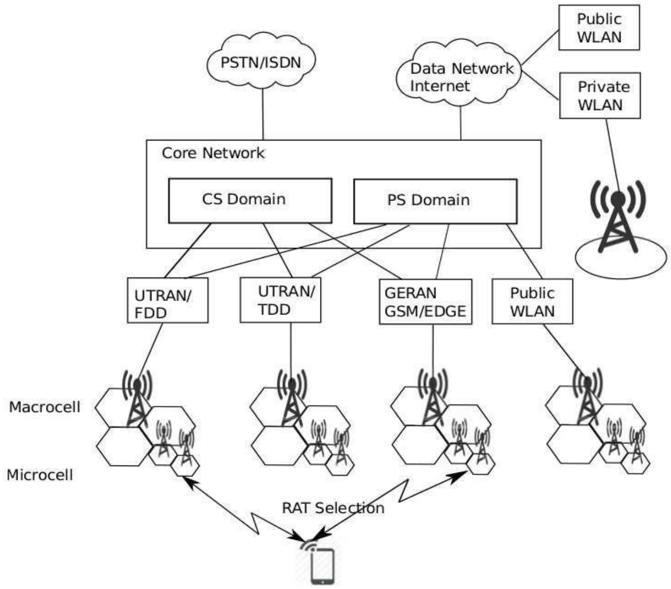


Fig. 1. Heterogeneous network environment

Mobile and wireless radio access networks differ in their radio coverage, air interference methods [2], access techniques, offered services, price [3] and ownership. It is worth to mention that we focus in our study just on common radio resource management in heterogeneous networks only for 3G technologies and earlier mobile and wireless access techniques. For systems and scenarios where different access technologies can be deployed and coordinate together is referred as Beyond 3G (B3G) systems (i.e. 4G, LTE, and WiMax) [4]; in order to achieve gain of such B3G networks, the available radio resources must be managed in a proper way. This trend introduces a new algorithms for managing the radio resources, that take into considerations the overall available resources offered by several RANs, such algorithms from the common perspective is called as Common Radio Resource Management (CRRM) algorithms, briefly the concept of CRRM uses a two-tier Radio Resource Management (RRM) model [5], including of RRM and CRRM entities as clarified in figure 2. Generally, Common Radio Resource Management (CRRM) involves a set of functions that are engineered to achieve a coordinated and efficient utilization of the available radio resources in complex scenarios that are including heterogeneous networks. More in details, CRRM policies should guarantee to meet the network operator's goals in terms of Quality of Service (QoS) and network coverage extension while increasing the overall capacity as high as possible.

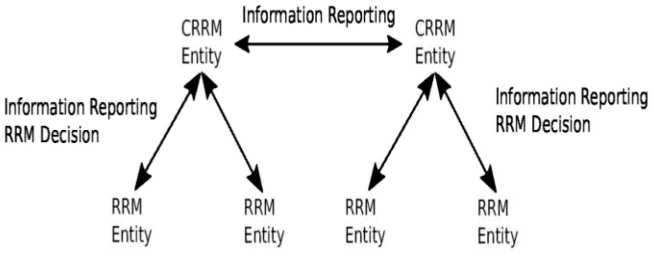


Fig. 2. CRRM functional model

The scope of this study is to analyze CRRM solutions with particular attention on implementations. Starting from an analysis of the state of the art, the most interesting solutions have been critically analyzed and then some in depth investigations on some of the identified solutions have been performed.

II. CRRM FUNCTIONALITIES

CRRM is designed to co-ordinately manage resources pools over the heterogeneous air interface in an efficient way. This efficiency depends on how to construct its functionalities. There exist a range of possibilities for the set of functionalities that CRRM entity may undertake, which mainly depend on the following two factors:

1) RRM or CRRM entity is the master to make radio resource management decisions.

2) The degree of interactions between RRM and CRRM entities

The RRM functionalities arising in the context of a single RAN are:

1) Admission control
2) Congestion control
3) Horizontal (intra-system) handover
4) Packet scheduling
5) Power control

When these functionalities are coordinated between different RANs in a heterogeneous scenario, they can be denoted as common (i.e. thus having common admission control, common congestion control, etc.) as long as algorithms take into account information about several RANs to make decisions. In turn, when a heterogeneous scenario is considered, a specific functionality arises, namely the RAT selection (i.e. the functionality devoted to decide to which RAT a given service request should be allocated). The functional model of CRRM is descriped and discussed in [6].

After the initial RAT selection decision [7],[8], taken at session initiation, vertical (inter-system) handover is the procedure that allows switching from one RAT to another for an on-going service. The successful execution of a seamless and fast vertical handover is essential for hiding to the user the underlying service enabling infrastructure see figure 3.

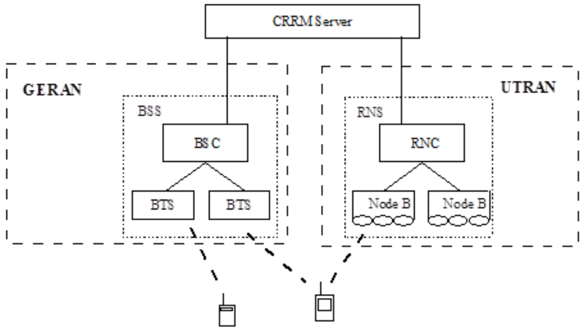


Fig. 3. Inter system handover between GERAN and UTRAN

Issues related to vertical handover comprise scanning procedures for the terminal to discover available RATs, measurement mecha- nisms to capture the status of the air interface in the different RATs, vertical handover triggers (i.e. the events occurring in the heterogeneous network scenario that require the system to consider whether a vertical handover is actually required or not), vertical handover algorithm (i.e. the criteria used to decide whether a vertical handover is to be performed or not) and protocol and architectural aspects to support handover execution.

A Markov model for performance evaluation of CRRM algorithms in a co-located GERAN/UTRAN/WLAN scenario is further discussed in [9]

III. RESEARCH METHODOLOGY

The methodology followed to carry out this research paper focused on the following steps:

1) Read and analyse accurately some research articles and technical available reports which report several CRRM solutions.

2) Investigate about the CRRM-algorithms and related work on the explicit subject, in order to derive the important features, requirements and architecture that can be used for implementing and modelling the CRRM-algorithms.

3) Using some software tools (models-simulator) al- ready developed in TiLab SpA [10] in order to model the behaviour of the system when CRRM solutions are applied; analytical models are based on Markovian Chain.

4) System modelization permits to analyse the QoS and system performance for the desired CRRM algorithm.

IV. RELATED WORK

Recently, different strategies of of Radio Resource Management (RRM) are independently implemented in each RAT. Since each RRM strategy [11] just only take into considerations the situations and conditions on only one RAT, thus none of the RRM strategy is suitable of heterogeneous networks. CRRM strategy, is also known as Joint RRM (JRRM) or, Multi-access RRM (MRRM), just strategies has be proposed in order to coordinate and optimize the utilization of different RATs. Many strategies has been proposed for CRRM, i.e. in [12] the results shows that CRRM has much better performance in networks in comparisons to that networks without CRRM, such performance gain is valid for networks with either real time (RT) and non-real time (NRT) services, in different terms, mainly capacity gain and blocking probability of the call [13].

The author in [6] proposed a Common RRM (CRRM) algorithm to jointly manage radio resources among different radio access technologies (RATs) in an optimized way. Moreover, a survey on the Common Radio Resource Management has been further analyzed in [14].

V. CRRM IMPLEMENTATION

The main factor for selecting suitable CRRM strategy and implementing its mechanism is depending on the functionalities associated to the CRRM, which determines and define the interaction between both RRM and CRRM entities. such interaction control is used for decision support and reporting the information between different network entities. It is important to note that, in all CRRM strategies, the trade-off between the any strategy gain and the typical network delay and signaling overhead must be considered.

Interworking architecture:

1) GERAN/UTRAN interworking: To establish a net- work connection between UTRAN and GERAN, both the Base Station Controller (BSC) and Radio Net- work Controller (RNC) must be connected to the same 3G CN, in particular to the Serving GPRS Support Node (SGSN) via the Iu interfaces (such interface is shown in 4).

2) 3GPP/WLAN internetworking: WLAN deploy- ments use a different network architecture from the architecture that is used by 3GPP system, whereas both UMTS and GSM/EDGE use 3GPP system net- working architecture. Thus, desired internetworking solution should consider both none technical and technical aspects. Thus, for supporting both CRRM and RRM functionalities, the APC (Access Point Controller), that is responsible for managing the radio resources utilized by the access points where the WLAN users are connected to, should be equipped with similar functionalities of the BSC and the RNC for both the GERN and UTRAN, as depicted in figure 4.

CRRM can be implemented as:

1) New separate node: CRRM entity can be imple- mented as a new separate node of the network (CRRM server). Furthermore, the CRRM server de- fines an open interface to facilitate internetworking between the CRRM node as well as the devices where RRM entities reside (i.e. APC, RNC and BSC). Such open internetworking interface is a common method generally is deployed in order to reduce or even remove the interoperability issues that are may introduce when different vendors components and equipments are interconnected. In most cases, such approach will boost both the cost and the time needed during any potential future upgrade tasks. More importantly, this approach will ensure that all the functionalities are centralised.

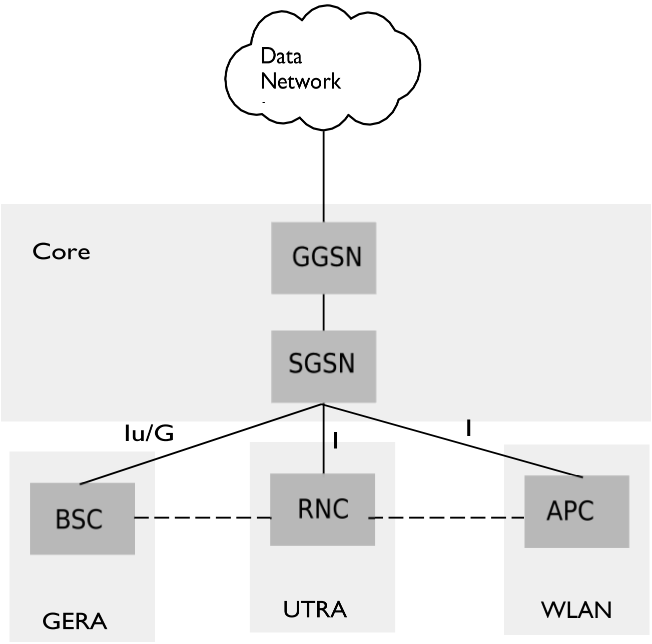


Fig. 4. WLAN/3GPP architecture

2) Integrate CRRM between existing nodes: CRRM functionalities can be integrated into existing nodes (integrated CRRM), in this case CRRM/RRM com- munications details not required to be defined in- priori and this detailed will depend on vendor im- plementation. The main advantage of this approach is that the system performance can be achieved with- out introducing additional delay, where the delay is important aspect especially for call setup, handover and channel switching.

VI. TECHNICAL KPI

Briefly we state the technical KPI for each of the earlier eight CRRM algorithms as follow:

CRRM-algorithm No-1 (Radio quality based inter-working mechanisms) Technical KPI: Capacity and performance increase of inter-working GERAN and UTRAN radio Access networks with respect to non inter-working case: lower blocking probability, lower out of coverage probability, lower radio link failures, higher throughput, etc.

CRRM-algorithm No-2: (CRRM perceived throughput) Technical KPI: Heterogeneous system throughput due to CRRM management

CRRM-algorithm No-3: (CRRM Cost Function) Technical KPI: CRRM KPIs: Delay, Blocking, Cost and Throughput

CRRM-algorithm No-4: (Coverage-based CRRM for Voice Traffic) Technical KPI: Voice Outage probability, blocking rate, and call dropping rate

TABLE I. OVERVIEW OF DIFFERENT CRRM SOLUTIONS

Short name	Category	Sub-category	Involved RATs	Desired Scenarios
Radio quality based inter-working mechanisms	CRRM	RRM strategies for VoWLAN and capacity estimation	UTRAN-R99, GERAN	Theoretical 2G-3G co-site
CRRM perceived throughput	CRRM	RRM strategies for com- bined usage of 2G, 3G and WLAN systems	GERAN, UTRAN R99 / R5 / R6, 802.11 a/b/g	Theoretical hot-spot urban
Coverage-based CRRM for Voice Traffic	CRRM	Coverage-based CRRM	Heterogeneous UMTS/GERAN	Realistic multi-floor building
CRRM Cost Function	CRRM	CRRM strategies for BS Selection	UMTS-R99, UMTS-R5 and WLAN	Theoretical hot-spot urban
MPLS based mobility man- agement and IP QoS	CRRM	Mobility management and QoS	LTE UTRAN	Specific defined scenario
Fittingness factor algorithm	CRRM	RAT selection strategies	UTRAN R99, HSDPA, HSUPA, GERAN, can be extended also to WLAN, etc.	Theoretical 2G-3G co-site
Common congestion control	CRRM	Inter-RAT RRM	UTRAN R99, GERAN	Theoretical 2G-3G co-site
Opportunistic CRRM	CRRM	RAT selection strategies	UTRAN R99, HSPA, and WLAN	Theoretical hot-spot main road

CRRM-algorithm No-5: (MPLS based mobility management and IP QoS) Technical KPI: From simulation results useful information can be derived concerning service degradation under different mobility managements scenarios. The simulations also show the performance improvement whilst using QoS routing (QOSPF) as opposed to normal routing (OSPF)

CRRM-algorithm No-6: (Fittingness factor algorithm) Technical KPI: KPIs depending on the service (e.g. delay for interactive users, error rate for conversational, etc.), total system throughput.

CRRM-algorithm No-7: (Common congestion control) Technical KPI: Packet delay, bit rate per service, load factor in UTRAN, reduction factor in GERAN.

CRRM-algorithm No-8: (Opportunistic CRRM) Technical KPI: Transmission delay. System capacity in the considered RATs

VII. OVERVIEW OF CRRM SOLUTIONS

The following group of CRRM algorithms come from IST-AROMA project [15]. These algorithms have been read and analysed carefully as an important task toward achieving the knowledge and requirements to model and implement CRRM algorithms and to evaluate them.

briefly, we provide a description about the studied CRRM algorithms.

1) Radio quality based inter-working mechanisms: 3GPP specified inter-working mechanisms between GERAN and UTRAN have been taken into account to identify useful CRRM strategies exclusively based on radio quality perceived users. Both inter-RAT cell re- selection and handover procedures were considered. In idle mode, simulation results show that inter-RAT cell re-selection can be used to implement different camping strategies. In connected mode, simulation results dealing with U2G (from UTRAN to GERAN) handover highlight that the handover procedure can be effectively exploited in order to take advantage of GERAN as a back-up system when the radio quality of UTRAN cell is not able to support users service (e.g. indoor users).

2) CRRM perceived throughput: Total data trans- mission delay times (connection setup, radio bearer establishment, TCP transmission etc) is used to cal- culate perceived user throughput for data transmis- sions. 2G, 3G and WLAN systems are analysed with different radio capabilities and with tight or loose WLAN coupling. Centralised CRRM algorithms are evaluated to analyse the total system throughput us- ing different radio access capabilities and different operator policies / CRRM algorithms.

3) CRRM Cost Function: An approach to integrate a set of KPIs into a single one, by using a Cost Function that takes a set of KPIs into account, providing a sin- gle evaluation parameter as output, and reflecting net- work conditions and CRRM strategies performance. The proposed model enables the implementation of different CRRM policies, by manipulating KPIs ac- cording to users or operators perspectives, allowing for a better QoS. Results show that different policies can in fact be established, with a different impact on the network.

4) Coverage-based CRRM for Voice Traffic: The coverage-based CRRM concept for hybrid FD/TDMA and CDMA cellular systems, which intend to improve system efficiency by taking advantage of the complementary characteristics of FD/TDMA and CDMA systems, i.e.

FD/TDMA is able to offer a rather static coverage and capacity while the coverage and capacity trade-off in CDMA is much more straightforward. This scheme has shown great potential to improve voice capacity in the heterogeneous environment.

5) Multiprotocol Label Switching (MPLS) based mobility management and IP QoS: A framework for QoS architecture with the MPLS-based micromobil- ity presented. The simulation platform includes the following functionalities: DiffServ and MPLS for the user-plane forwarding, QoS-enabled Open Shortest Path First (QOSPF) for the routing, bandwidth broker for the resource reservation and admission control and IP micro-mobility for the intra-domain mobility management.

6) Fittingness factor algorithm: It consists in a new generic framework for developing CRRM strategies in heterogeneous scenarios was presented. It captures the different degrees of heterogeneities that can be found in the network (including RAT and terminal capabilities as well as the suitability of one or an- other RAT depending on the current interference, path loss and load conditions) by means of the so- called fittingness factor of one cell in one RAT. From this metric, new RAT selection schemes both at the session initiation and during the connection lifetime have been defined.

7) Common congestion control: This algorithm addresses how to solve congestion situations in UTRAN/GERAN networks by means of executing vertical handover procedures between both RATs and RAT-specific procedures, like bit rate reduction in UTRAN.

8) Opportunistic CRRM: Opportunistic CRRM is intended for services without stringent delay con- straints. It is based on the concept that these services allow waiting until the coverage area of a high speed RAT (e.g. WLAN, High Speed Packet Access (HSPA), etc.) is found instead of making use of RATs with continuous coverage (e.g. GERAN, UMTS) with a more reduced bit rate.

In Table I, we present a brief comparison between the earlier algorithms under study in terms of Short name, Category, Sub-category, Involved RATs and Desired Scenarios.

VIII. Considerations on Techno-Economic Impacts of CRRM Solutions Envisaged by IST-Aroma Project

The previous eight RRM/CRRM algorithms are already identified within the legacy IST-AROMA project and assessed only from the technical point of view. On the other hand we try in this section to describe the potential economic advantage of using some of these RRM/CRRM algorithms.

A Cost Function for Heterogeneous Networks Performance Evaluation Based on Different Perspectives is discussed and described in [16] Each CRRM algorithm can be evaluated from the techno-economic point of view on the basis of the effects produced on the system. The main evident effects of the reported CRRM algorithms that have techno-economic implications are the QoS improvement either for voice users or data users or both and the increase of the network capacity. As a matter of the fact, both the QoS improvement and the capacity increase can reduce effectively network CAPEX and OPEX. It is worth noting that these factors are also able to increase directly the operators revenues.

In order to evaluate the economic impact of the algorithms, specific economic models should be proposed .In some algorithms, which improve the QoS, we can notice that, the objective of the network operator is to support its customer with the required QoS in profitable way to derive additional revenues. As example, it is possible to divide the users into two profiles. Each profile depends on the sensitivity to price and reflects the willingness of pay of the user:

1) Flat (consumer) profile: it is naturally the cheapest option for all users.

2) Business profile: users pay extra than flat profile.

Clearly, in order to encourage users to move towards the more expensive option, the operator should guarantee the perceived QoS according to the contracted QoS.

On other hand where the algorithms affect the system capacity, we can notice that increasing the capacity can increase also the revenue for operator. Within the context of the techno- economic analysis, it is also possible to compare the revenue achievable by means of the specific CRRM algorithm under investigation in term of capacity increase, and the cost for achieving the same capacity increase using new additional sites and network resources.

The approaches summarized here have been followed during the internship in order to approach a qualitative evaluation of the techno-economic impacts of the CRRM algorithms.This job has been accomplished by collecting some relevant figures and information concerning cost of network equipments and market trends. Taking into account all the previous considerations the work can be useful to derive useful information for supporting strategic and economic-driven decisionsconcerning the exploitation of the considered RATs within the context of a heterogeneous network. This kind of results could be valuable not only from the research point of view but also by network operators dealing with economic evaluations of the impacts of the deployment of new technologies and apparatus in order to improve the network QoS and to extend the capacity of the mobile access network.

IX. Conclusions and Future Work

This study describes the existing heterogeneous network scenario consisting in a mix of different RATs. in the telecommunications field. This scenario arises the need for network functionalities (i.e. CRRM solutions) devoted to mange the pool of resources offered by different RATs to get benefit of the specific characteristics of each RAT with the aim of increasing the network capacity and improving the QoS. While the proper selection of CRRM- algorithm reflect the gain that can be obtained from the heterogeneous scenarios. This Study provides a wide vision of actual scenario of the mobile communications, which is going towards a heterogeneous network where the development and improvement of CRRM algorithms is going to be the challenge.

References

[1] M. Gerasimenko, D. Moltchanov, R. Florea, S. Andreev, Y. Koucheryavy, N. Himayat, S.-P. Yeh, and S. Talwar, “Cooperative radio resource management in heterogeneous cloud radio access networks,” *Access, IEEE*, vol. 3, pp. 397–406, 2015.

[2] V. Pauli, J. D. Naranjo, and E. Seidel, “Heterogeneous lte networks and inter-cell interference coordination,” *Nomor Research GmBH*, pp. 1–9, 2010.

[3] S. Farhat, A. E. Samhat, S. Lahoud, and B. Cousin, “Pricing strategies in multi-operator heterogeneous wireless networks,” 2015.

[4] R. Agust´ı, “Radio resource management in beyond 3g systems,” in *Electrotechnical Conference, 2006. MELECON 2006. IEEE Mediterranean*. IEEE, 2006, pp. 569–574.

[5] N. Passas, S. Paskalis, A. Kaloxylos, F. Bader, R. Narcisi, E. Tsontsis, S. Jahan, H. Aghvami, M. O’Droma, and I. Ganchev, “Enabling technologies for the always best connectedconcept,” *Wireless Communications and Mobile Computing*, vol. 6, no. 4, pp. 523–540, 2006.

[6] J. Pe´rez-Romero, O. Sallent, R. Agust´ı, P. Karlsson, A. Barbaresi, L. Wang, F. Casadevall, M. Dohler, H. Gonzalez, and F. C. Pinto, “Common radio resource management: functional models and implementation requirements.” in *PIMRC*, 2005, pp. 2067–2071.

[7] J. Pe´rez-Romero, O. Sallent, and R. Agust´ı, “Policy-based initial rat selection algorithms in heterogeneous networks,” 2005.

[8] H. M. ElBadawy, “Optimal rat selection algorithm through common radio resource management in heterogeneous wireless networks,” pp. 1–9, 2011.

[9] L. Wu, K. Sandrasegaran, and M. Elkashlan, “A markov model for performance evaluation of crrm algorithms in a co-located geran/utran/wlan scenario,” pp. 1–6, 2010.

[10] T. SpA, “TiLab SpA Website,” http://www.telecomitalia.com/tit/en/innovazione/i-luoghi-della-ricerca/tilab.html, 2007, [Online; accessed 01-April-2015].

[11] S. Farhat, A. Samhat, S. Lahoud, and B. Cousin, “Best operator policy in a heterogeneous wireless network,” in *e-Technologies and Networks for Development (ICeND), 2014 Third International Conference on*, April 2014, pp. 53–57.

[12] A. Tolli, P. Hakalin, and H. Holma, “Performance evaluation of common radio resource management (crrm),” in *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 5, 2002, pp. 3429–3433 vol.5.

[13] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, “Recent advances in radio resource management for heterogeneous lte/lte-a networks,” *Communications Surveys Tutorials, IEEE*, vol. 16, no. 4, pp. 2142–2180, Fourthquarter 2014.

[14] L. Wu and K. Sandrasegaran, “A survey on common radio resource management,” in Wireless Broadband and Ultra Wideband Communications, 2007. AusWireless 2007. The 2nd International Conference on, Aug 2007, pp. 66–66.

[15] A. I. Project, “AROMA IST Project Website,” http://www.aroma-ist.upc.edu/, 2007, [Online; accessed 01-April-2015].

[16] A. Serrador and L. Correia, “A cost function for heterogeneous networks performance evaluation based on different perspectives,” in *Mobile and Wireless Communications Summit, 2007. 16th IST*, July 2007, pp. 1–5.

Big-Learn: Towards a Tool Based on Big Data to Improve Research in an E-Learning Environment

Karim Aoulad Abdelouarit
Laboratory of LIROSA
Faculty of Sciences
Tétouan, Morocco

Boubker Sbihi
Laboratory of LIROSA
Ecole des Sciences de l'Information
Tétouan, Morocco

Noura Aknin
Laboratory of LIROSA
Faculty of Sciences
Tétouan, Morocco

***Abstract*—In the area of data management for information system and especially at the level of e-learning platforms, the Big Data phenomenon makes the data difficult to deal with standard database or information management tools. Indeed, for educational purposes and especially in a distance training or online research, the learner that uses the e-learning platform is left with a heterogeneous set of data such as files of all kinds, curves, course materials, quizzes, etc. This requires a specialized fusion system to combine the variety of data and improve the performance, robustness, flexibility, consistency and scalability, so that they can provide the best result to the learner The user of the e-learning platform. In this context, it is proposed to develop a tool called "Big-Learn" based on a technique to integrate the mixing of structured and unstructured data in one data layer, and, in order to facilitate access more optimal search relevance with adequate and consistent results according to the expectations of the learner. The methodology adopted will consist initially in a quantitative and qualitative study of the variety of data and their typology, followed by a detailed analysis of the structure and harmonization of the data to finally find a fictional model for their treatment. This conceptual work will be crowned with a working prototype as a tool achieved with UML and Java technology.**

***Keywords*—big data; e-learning; data structuring; learning; digital pedagogy**

I. Introduction

The users of the Internet, whether it was humans, programs or services provide every day enormous amount of data that become so difficult to manage and deal with traditional database management tools. The massive exploitation of such information data herein is called Big Data or massive volumes of data.

Moreover, with the emergence of Web 2.0, a new vision of the Web put the user at the center of information considering it as a potential producer of web content and not just a consumer [8]. This radical change has significantly increased the amount of information [10]. Consequently, the proliferation of types of information from multiple sources such as social networking, services, blogs, information aggregation websites, videos, images, text, creates a wide variety of data types beyond traditional relational data. These data do not exist in a perfectly ordered form and are not amenable to analytical operations. They no longer fall within the net structures, easy to consume, rather they are semi-structured or unstructured. This is within the aspect range of data which represent the second V in the design of the Big Data phenomenon. And consequently, this leads to ask: how it is possible to deal with and process these unstructured data to make them consumable by human users and / or applications?

Indeed, it is possible to find this problematic of data variety on educational purposes, and especially during a online search or an e-learning process. The learner that uses the e-learning platform is left with a mix of data that do not meet necessarily their expectations and they sometimes even prove useless against expected results. It is in this context that it is proposed to develop a methodology based on a tool called "Big-Learn" to integrate the mix of structured and unstructured data in one data layer to facilitate access and more optimal relevance search with adequate and consistent results according to the expectations of the learner. So, is there a fictitious model to represent and process this type of data that are not necessarily text? Also, is the semi-structured databases NoSQL provide enough structure to organize the unstructured data? knowing they do not require an exact schema data before storing.

The following paragraph present the state of the art concerning Big Data in the e-learning environment and the describtion of the problematic of data variety and their impact on the educational purposes ; the paragraph 3 expose the concepts and approach of the Big-Learn tool using the online search as case study. The last paragraph present a general conclusion putting forward a series of perspectives.

II. Big Data in E-Learning Environment

A. Definition and state of the art

The society has experienced in recent years the arrival of Big Data phenomenon, So much data is available and requires powerful computers and algorithms process them. Attal Butte estimates the volume of data produced each year to four-zetta bytes (1 byte = zetta-10 to the power 21 bytes) [1]. Currently, 2.5 terabytes of data are produced every day in the world. By the year 2020 it is estimated that the data size will be multiplied by 50. Google receives 40,000 requests for information every second, 72 movies are set to YouTube every minute and 217 new Smartphone users are counted every minute. [7] Today the information is coming from all sides: geolocation sensors, data from smartphones (connection logs, appeals, etc.), data posted on social networks, video and satellite images, Transactions customers, sensor forms of movement or connected objects, etc. Concretely, this is the real-time development of a large mass of data that goes far beyond the capacity of conventional processing and analysis tools (relational databases, SQL, etc.). These mass data needs to be analyzed and processed for their use and consumption on the part of users and applications.

However, a number of authors [2] [6] postulate some provocations on the Big Data phenomenon that require critical thinking. The increasing use of big data questions some assumptions about traditional knowledge in the context of a claim of Big Data to "objectivity and accuracy [who] are deceiving." It is also necessary to realize that "all the data is not equivalent" [2]. To better explain this, the Table 1 summarize some advantages and limits of Big Data phenomenon.

TABLE I. SOME ADVANTAGES AND LIMITS OF BIG DATA

Advantages	Limits
• The ability to search and cross massive data sets; • Completeness of the perimeter and the capture of entire populations of the systems; • Targeting maximum detail, aiming the common fields for the combination of data sets; • The flexibility and extension: Easily add new fields; • Scalability: the potential to grow rapidly; • The prediction of future performance and identify potential problems; • The interpretation of the actual operational data; • The evaluation of the performance of the organization or institution; • Decision making thanks to the researched information at the right time.	• The creation of new fractures: increasing inequalities and injustices that exist. • The complexity of managing and analyzing large amount of heterogeneous data. • Uncertainty of informations coming from different sources; • The loss of value and / or reliability of data coming from different systems. • Problems related to the ethics and confidentiality of information; • Lack of regulations against abuse and bad data uses. • Misinterpretation of information; • Collecting incorrect data that can lead to erroneous results.

As it is shown, the Big Data phenomenon offers several benefits for the completeness and flexibility of data use that themselves can cause bad consequences on the objectivity and accuracy of the interpreted information. The analysis and exploration of big data exceeds human capacity, which requires the use of computer systems powerful and able to explore them. But these data does not occur in an ordered form and are not ready for analytical operations, they rather come in a semi-structured or unstructured form. It is this dimension that concern the subject of this article and especially the problematic of the non structured types of data produced by Big Data.

The Big Data phenomenon has obviously impacted the learning environment and the distance training. It has facilitated the creation of a mixture full of learning opportunities and allows the learners to improve their training practices and experiment with open educational resources, especially the massive and open online courses (MOOC) and the distance learning via e-learning platforms. With the emerging technologies in the Web, access to information has become easier with the ability to work and learn effectively, regardless of educational structures that have been the norm for centuries [3] [4]. This put in place new structures and new working environments, enabling independent learning, but that does not mean that everyone is able to do it effectively [5]. Two major factors are the basis for the study of learning in a massive and open online environment: learner's autonomy and the quality of the massive submitted information. It is this last factor that concern the subject of this article.

B. *Problematic of variety and non-structuring massive data and their impact on the educational purposes*

Big Data knows several challenges and opportunities as well as involvement of several technologies because of the flood of data produced each year by users and companies. The information of the Big Data comes from many sources of data. The Table 2 shows some examples of these sources like the Web, the Internet, communication objects, genomic sciences, astronomy, commerce and public data.

TABLE II. SOME DATA SOURCES OF BIG DATA

Data source	Examples
Web	Access logs, social networks, e-commerce, documents, photos, videos, etc. (example: Google treated 24 petabytes of data per day with MapReduce in 2009).
Internet and communication objects	RFID, sensor networks, telephony call logs.
Genomic science, astronomy, subatomic physics, climatology, etc.	CERN announces generate 15 petabytes of data per year with the LHC. The German research center on climate manages 60 petabytes of database.
Marketing	The transaction history in a hypermarket chain.
Public	Open Data

As presented in the table, the integration of data in the Big Data processes covered several unstructured data (sensor data, web logs, RFID, social networks, documents).

It is possible to find the problem of the data variety for educational purposes especially in online search or distance training. The learner that uses the e-learning platform is left with a mixture of data such as files of all kinds, curves, course materials, quizzes, etc. Also, in the educational component, the social tools of web 2.0 allows to create and publish any type of educational content such as lectures, exercises, assignments or bibliographies and digital resources enable an informal collaborative learning known as "Learning 2.0"[9]. However, these data do not always meet the needs of the learner and they sometimes even prove useless against expected results. A new challenge is then born in data analysis, it is to make significant progress in the process of this type of unstructured data which the amount is continuously growing. To do that, the use of specific IT tools for the processing of unstructured data has become essential. Knowing that relational databases are not always the best solution because of their static patterns.

III. TOWARDS A BIG DATA SYSTEM FOR PROCESSING MASSIVE DATA IN THE E-LEARNING ENVIRONMENT

A. *Case study: the use of the online search*

1) Context

Since becoming aware of innovations in learning, technologists have begun to design and develop tools to help learners to better understand this new way of teaching connected to data in constant evolution. To success the purpose of the e-learning, it is crucial to create a trusted environment where learners feel comfortable. A place that can aggregate content and imagine it as a community where dialogue flows and interactions and content can be simple to use. This will enable learners to develop clear ideas and evolve in their learning in depth.

In this context, it is proposed to create an educational platform that would support learners in their environment. Research involving the design and development of this platform is working in many directions, but here the object is to report some progress in education, advances on issues concerning self-learning and online search in a massive data environment. And to better understand the study, the choosen case is the use of the online search by learners who seek to acquire information about a specific course or theme given for the purpose of learning and documentation, and, to see if there are other extra dimensions that could be added following the study on learning in a massive data environment.

The results of the data analysis for the online search scenario will allow to delineate the context of the future system and to better understand the design of a methodology based on the tool that will integrate the mixing of structured and unstructured data in one layer to facilitate access in addition to an optimal search relevance with adequate and consistent results according to the needs of the learner. The Fig. 1 shows the Big-Learn system usage scenario illustrated by a sequence diagram.

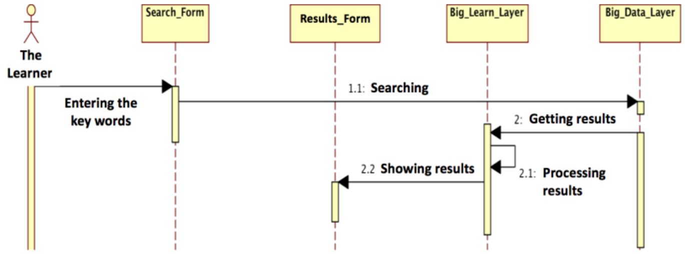


Fig. 1. The sequence diagram of the general use for the Big-Learn system

The user of the system (the learner) accesses the search interface to enter the keywords of his information request. These keywords are sent to the system for retrieving information corresponding to their semantics from the Big Data layer. Then, the results data is processed at the Big-learn layer to structure, classify and send the processed data to the results page of the system, which will be displayed to the learner, the user of the Big-Learn system.

2) *Methodology*

To a better design and development of the futur system, it is necessary to study the circumstances of the learning that takes place on online networks and distance training. It is important to know the relevance of the learning experience of people in online networks in which they find the information they are likely to consume. As part of this study, informational or learning data are defined as data collected from online open spaces where people access remotely, while communicating with others via blogs, audiovisual, wikis and other sources of information and other remote communication resources. Constraints and challenges emanating from such an environment show themselves well in problems related to the study of human behavior, as well as other constraints involved, including the variability of the network and data, power relations on the network size and the generated content. Relevant analysis requires an approach by mixed methods and results in a new ethics and questions about the confidentiality of information or data.

For this study, it is proposed to conduct a sampling survey concerning the criteria of the future system and indicators of performance and learning quality in the use of online search, through the design and the submission of survey to a set of students who represent the future users of the online search and distance training platform. The content of the survey describe the different factors suceptibles to impact and affect the learner when using the online search, and the criteria of simplicity and ergonomics of the use of the future solution.

The questions of the survey were raised in relation to the aspect of the presentation and quality of information in the online search results in a documentation or on a distance training course or theme. The classification of information also plays a dominant role in the organization and fair presentation of the results of the search. The relevancy of the result and the content is crucial to the learner using online search. Thus, several elements must be examined and redefined to design the right solution.

The Google Forms tool will be used to design and submit the survey. It is a tool to plan events, to conduct a survey or poll to subject students to a questionnaire or to easily collect information online. Forms can be created from Google Drive or an existing spreadsheet can collect answers to questions on the form. The target audience consists of two samples, the first few groups of students from the ESI (Science School of Information) of Rabat, and the second is a group of students from the FS (Faculty of Science) of Tetouan.

The goal of this investigation is to identify the key elements and factors that may impact and influence the environment when the learner uses the online search for learning purposes and documentation on a given theme or topic, and to enable to deduct thereafter performance and functioning indicators related to learning in the Big Data environment to better design and implement an adequate system and meeting the expectations of most learners. Table 3 highlights some examples of factors to consider elements during this investigation.

TABLE III. THE FACTORS ELEMENTS OF THE ONLINE SEARCH SURVEY

Data source	Examples
The type of search result that the learner prioritizes at its online search	video, document, image, article, ...
The criteria of a relevant and fractueuse search	Number of views, date of publication, source of the element, ...
The number of results that the learner prefer by page	5, 10, 50, ...
The response time in seconds that the search query take to be sent	1s, 5s, 10s, ...

As described in this table, several factors can influence the process of using the online search by the learner. Thus, the results of its application will depend on the relevance and proper configuration of these criteria.

B. The Big-Learn System

It is important to identify a sophisticated strategy to combine different types of data in a way that they provide the best result to the learner, the user of the e-learning platform. In this context, it is proposed to develop a technique based on a tool called "Big-Learn" that integrates the mix of structured and unstructured data in one data layer to facilitate access in addition to an optimal relevance of search with adequate and consistent results according to the expectations of the learner. The adopted method will consist initially in a quantitative and a qualitative study of the variety of data and their typology, followed by a detailed analysis of the structure and harmonization of the data to finally find a fictional model for treatment of such data. This conceptual work will be crowned with a working prototype as a tool achieved with UML and Java technology. Fig. 2 shows the functional architecture of the Big-Learn system.

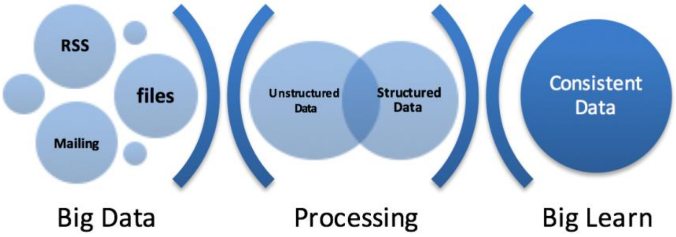


Fig. 2. The functional architecture of the Big-Learn system

As is described, it is proposed to develop a top layer to the raw and varied data coming from "Big Data" providing thereafter the storage, retrieval and dissemination of consistent information to research information requested by the user of the Big-Learn platform.

The scenario of the use of the online search that has been already mentioned above, can be achieved through the application components of the Big-Learn system as described in Fig. 3.

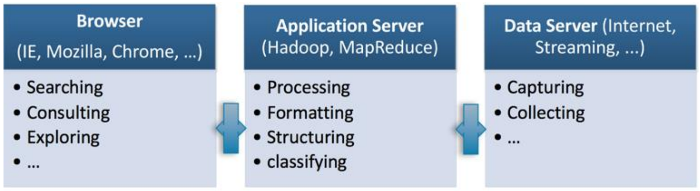


Fig. 3. The application architecture of the Big-Learn system

Thus, the data server will handle the capture and collection of massive data (Big Data) and then, the application server can carry out the treatment, structuring, formatting and classification required for these raw data and thus make them consumable by the presentation layer at the end and that will be accessible via the web browser by the user of the Big-Learn system.

The interface of the tool will be an easy space to use for the learner to make online search for learning or documentation on a specific subject or theme, the example shown in Fig. 4 shows the use of the system via its interface to search information about the subject of the use of the system "Viber": The application for free calls and messages.

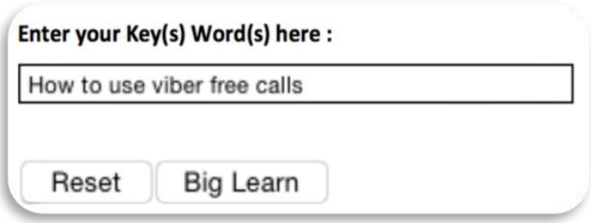

Enter your Key(s) Word(s) here : How to use viber free calls

Fig. 4. The user interface of the Big-Learn system

After entering the keywords to search the specific thematic as shown in the previous figure, the Big-Learn system performs the capture of all types of data (text, image, video, audio, etc.) related to the subject of the theme and group them in its raw data layer as shown in Fig. 5. It then includes data of any type, such as posts, pictures, videos, audio tracks, etc.

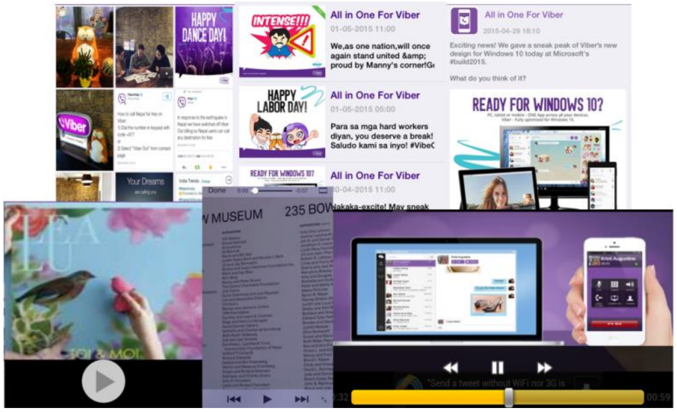


Fig. 5. The raw data collected related to the requested theme

The big-Learn system thereafter proceed with the treatment of the raw data to make it consumable by the user, via classification, structuring and formatting of these data at the presentation layer, thus allowing an organized display and ergonomics at the user interface of the system, as illustrated at Fig. 6.

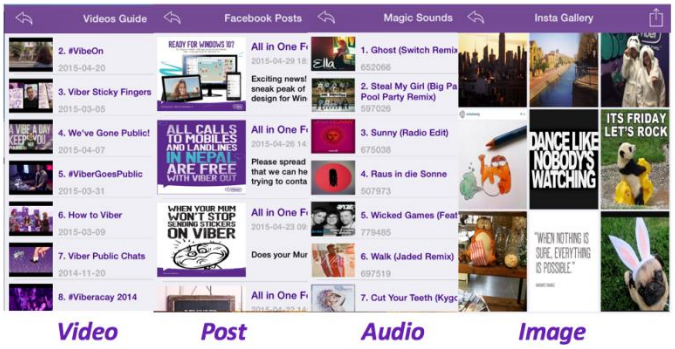


Fig. 6. The organization of results data according to their typology

As it shown, the data related to the video type classified in the same category even though they come from different sources (youtube, vimeo, dailymotion, ..), also, for the type of post data (facebook, twitter, google +, ...), audio data type (soundcloud, deezer, ...) and finally the type of the image data (instagram, picassa ...).

IV. Conclusion and Future Work

The available tools actually on the market make it possible either to analyze structured or unstructured data, but not both at the same time. Consequently, little Big Data technologies provide integration of various types of data and align with the structured data. It is in this context that it is proposed to develop a methodology based on a tool to integrate the mix of structured and unstructured data in one data layer to facilitate access and more optimal relevant search with adequate and consistent results that meets the expectations of the learner. The solution will also enable the detection of language elements, turn them into a data type that can be manipulated and be the object of the processing of the consumption of the information. The adopted method will consist initially in the study of the criteria and factors impacting the environment of the learner towards massive data offered by the Big Data via the case study of using the online search for learning purposes or documentation on a given theme. This, through the creation and submission of the survey corresponding to a sample of learners using online search for their learning. This will be followed by a detailed analysis of the results collected from these survey and that will frame the functional and technical requirements of the future solution to finally design a hypothetical model for the treatment of these heterogeneous mass of data.

References

[1] Attal Butte. Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? (2014). Adult Education, 2014, vol. 26, no 1, p. 35-55.

[2] Boyd, D., & Crawford, K. (2013). Six provocations for Big Data.

[3] Downes, S. (2010, May 12). The role of the educator.

[4] Fournier, H., & Kop, R. (2011) Factors affecting the design and development of a Personal Learning Environment: Research on super-users, in the International Journal of Virtual and Personal Learning Environments, 2 (4), 12-22.

[5] Kop et Bouchard, 2011 Kop, R., & Bouchard, P. (2011). The role of adult educators in the age of social media. In M. Thomas (Ed.), Digital education:Opportunities for social collaboration (pp. 61-80). New York, NY:Palgrave Macmillan.

[6] Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society*, July-September pp. 1—13. DOI: 10.1177/20253951714541861.

[7] Miranda, S. (2013). « De Big Brother au Big Data », Conférence de Big Data, Université Sophia Antipolis.

[8] O'Reilly, T. 2005. "What Is Web 2.0, Design Patterns and Business Models for the Next Generation of Software",

[9] SBIHI, Boubker, EL KADIRI, Kamal, et AKNIN, Noura. Towards an Implementation of the Concepts of E-Learning 2.5 through one Group of ten Master’s Learners: Case of the UML Course. *International Journal of Emerging Technologies in Learning (iJET)*, 2013, vol. 8, no 4, p. 68-73.

[10] SBIHI, Boubker et KADIRI, Kamal Eddine El. Towards a participatory E-learning 2.0 A new E-learning focused on learners and validation of the content. *arXiv preprint arXiv:1001.4738*, 2010.

Classifying three Communities of Assam Based on Anthropometric Characteristics using R Programming

Sadiq Hussain
System Administrator
Examination Branch
Dibrugarh University

Runjun Patir
Department of Anthropology
Dibrugarh University
Dibrugarh

Prof. Jiten Hazarika
Department of Statistics
Dibrugarh University
Dibrugarh

Dali Dutta
Department of Anthropology
Dibrugarh University
Dibrugarh

Prof. Sarthak Sengupta
Department of Anthropology
Dibrugarh University
Dibrugarh

Prof. G.C. Hazarika
Department of Mathematics
Dibrugarh University
Dibrugarh

***Abstract*—The study of anthropometric characteristics of different communities plays an important role in design, ergonomics and architecture. As the change of life style, nutrition and ethnic composition of different communities leads to obesity epidemic etc. The authors performed two experiments. In the first experiment, the authors tried to classify three communities of Assam, India based on anthropometric characteristics using R Programming. The authors mined out the statistically significant anthropometric characteristics among the Chutia, Mising and Deori communities of Assam. In the second experiment, the authors performed the Cochran Mantel Haenszel test to find out the association between the communities and BMI based nutritional status stratified by the age of the people studied.**

***Keywords*—Data Mining; Classification; R Programming; Logistic Regression; Cochran Mantel Haenszel Test**

I. Introduction

The measurement of human individual termed as anthropometry plays a crucial role in designs and architecture where the statistical data about the anthropometric characteristics are used to optimize products. The need of regular updating of anthropometric characteristics increases because of the changes in life style, nutrition etc. among different communities lead to changes in distribution in body dimensions.

Physical Anthropology is mostly concerned with the taxonomic classification of human population at both micro and macro level to understand the process of human evolution in space and time. As such it deals with the phylogenetic position of human populations in terms of their differences and similarities mainly in respect of morphological and anthropometric characters. One of the natural assets of peoples belonging to different population groups is their body build or physique. This can be measured and varied.

The difference or dissimilarities between generations within a population or between population within a major ethnic groups in respect of anthropometric and genetic traits are considered as the ongoing process of human evolution and is subject to a number of evolutionary forces which act differently in different population.

This work attempts to classify three communities –Chutiya, Deori and Mising of Assam based on anthropometric characteristics.

The Chutiya, one of the numerically dominant Other Backward Communities (OBC) of Assam, form one of the old populations of Assam. The Chutiyas are confined to Dibrugarh, Sibsagar, Jorhat, Golaghat, Lakhimpur and Dhemaji districts of Assam. These districts are called upper Assam districts. The Chutiyas had their own kingdom in upper Assam region. The Ahom, a Tai Mongoloid population came to Assam in the 13th Century and the Chutiyas tried to resist their aggression. In the long run, the Ahom overrun the Chutiya kingdom. Linguistically, the Chutiyas belong to Tibeto-Burman family. However, they accepted Assamese Language and they are Indo Mongoloid. The Chutiyas may be subdivided into several groups like Hindu-Chutiya, Ahom-Chutiya, Deori-Chutiya, Borahi-Chutiya etc. The Chutiyas are by religion Hindu.

The Deori were traditionally engaged in priestly activities of the Chutiyas. They were one of the major sub-divisions of the Chutiya [3]. The word Deori means in-charge of a temple or the priest. Nowadays, they however like to identify themselves as ‘Gimasaya’,meaning ‘the children of the Sun and Moon’.The Deori are the Tibeto-Mongoloid tribal groups of Assam. They are recognized as one of the plain scheduled tribes of Assam. According to 2001 Census, the total population of Deori is 41,161 in Assam. The original habitat of the Deori was in the Lohit district of Arunachal Pradesh. They migrated to Brahmaputra valley, Assam to escape from frequent troubles created by the Mishmis and the Adis.

The Mising are another Indo-Mongoloid Schedule Tribe of Assam. The Mising is synonymous with Miri, which means mediator, intermediary, interpreter.[3]. According to Census of 2001, the population of Mising is estimated at 5,87,310. The Misings were inhabitants of the hilly ranges that lie between the Subansiri and the Siyang districts of Arunachal Pradesh. They migrated down to the plains of Assam from an area upstream of the Dihong river in search of better economic life before the advent of the Ahom rules in Assam. Since then the Misings have been living mostly along banks of Brahmaputra

River and its tributaries. The Mising still speak their own dialect, which is akin to that of Adis of Arunachal Pradesh and possess their traditional ways of living. Originally, they were worshiper of Donyi (Sun) and Polo (Moon), but at present some of them are followers of Mahapurushia Vaishnav Dharma propounded by Srimanta Sankardeva during 15th and 16th centuries A.D.

In the present study, the authors made an attempt to understand the phylogenetic position of the populations under investigation. All the populations considered represent numerically small endogamous Mendelian population. In the present day context, however, due to globalization there are possibilities of bio-cultural disintegration in these populations due to increasing contact with relatively advanced neighbouring peoples.The authors analyzed anthropometric characteristics of above three communities viz. Chutia, Mising, and Deori of Assam and tried to classify them using the Logistic Regression Model. The Mantel-Haenszel odds ratio was also calculated to find out the association between the communities and BMI based nutritional status stratified by the age of the people studied.

II. Literature Review

In this section, some of the important literatures related to present study particularly related to methodological aspects, are briefly discussed.

Cancer classification and prediction is an important research area and it is one of the most important applications of DNA microarray due to its potentials in cancer diagnostic. The logistic regression model is used successfully for cancer classification and prediction. [14]

Logistic regression analysis can also be used in text mining poses computational and statistical challenges. Genkin et. al. [4] used Bayesian logistic regression approach that uses a Laplace prior to avoid over fitting and produces sparse predictive models for text data. They used it for document classification problems.

Using microarray data, Logistic regression may also be used for disease classification. Liao et. al. [3] proposed a parametric bootstrap model for more accurate estimation of the prediction error that is tailored to the microarray data by borrowing from the extensive research in identifying differentially expressed genes.

A comparative study between the linear discriminant analysis and logistic regression may also be found in [10]. They consider the problem of choosing between the two methods, and set some guidelines for proper choice. The comparison between the methods was based on several measures of predictive accuracy. The performance of the methods was studied by simulations.

Another comparison of discriminant analysis and logistic regression were made in [9] using two data sets from a study on predictors of coliform mastitis in dairy cows. Both techniques selected the same set of variables as important predictors and were of nearly equal value in classifying cows as having, or not having mastitis. The logistic regression model made fewer classification errors. The magnitudes of the effects were considerably different for some variables.

Logistic regression may also be used for classification of community based on Anthropometric predictors. In [12], it was aimed to examine the associations of anthropometric indices with gestational hypertensive disorders (GHD), and to determine the index that can best predict the risk of this condition occurring during pregnancy among Australian aboriginal women. The associations of the baseline anthropometric measurements with GHD were assessed using conditional logistic regression.

Kitamura et al. [6] used Hierarchical logistic regression to find out association between delayed bedtime and sleep-related problems among community dwelling 2-year-old children in Japan. It was carried out with the incidence of each sleep-related problem (present two or more times per week) as the dependent variable and bed time as the independent variables in model.

Cephalic Index (CI) is used in classifying the racial and gender differences. Lobo et al. [8] used CI for classifying Gurung Community of Nepal based on anthropometric indices.

III. Anthropometric Characteristics

There are various anthropometric characteristics based on which different communities can be classified. In this section, the anthropometric characteristics considered in our present study are briefly discussed.

Weight: It is taken by means of standard portable calibrated spring weighing machine. The individual is asked to stand at the centre of weighing machine with minimal clothing, looking straight and breathing normally. Body weight of each subject measured to the nearest 0.1 kg on the weighing machine with minimum cloths adjustment. The weighing machine is checked from time to time with a known standard weight. No deduction is made for the weight of light apparel while taking the final reading.

Stature: It measures the vertical distance between the floor and the vertex. While taking the stature, the subject is asked to remove shoes and stand erect against a wall with the heels, buttocks and shoulders and back of the head touching the wall and the head resting on the Frankfurt Horizontal Plane. The anthropometer is placed at the back and between the heels of the subject, taking care that it is kept vertical. The sliding sleeve of the anthropometer is then lowered down towards the middle of the head so that it could touch the vertex. Reading in centimeter and its fraction were recorded.

Blood pressure: Blood pressure or arterial blood pressure is the pressure exerted by circulating blood upon the walls of the blood vessels. During each heart-beat, blood pressure varies between a maximum (systolic) and a minimum (diastolic) pressure.

A person’s blood pressure is usually expressed in terms of the systolic pressure over diastolic pressure and is measured millimeters of mercury (mmHg). The subjects were classified following the American Medical Association [1].

TABLE I. CLASSIFICATION OF BLOOD PRESSURE

NORMOTENSIVE (mmHg)	SBP < 120 and DBP < 80
PRE-HYPERTENSIVE (mmHg)	SBP 120 – 139 ; DBP 80 – 89
HYPERTENSIVE (mmHg)	SBP > 140 ; DBP > 90 - 99

Sitting height: It measures the vertical distance from the vertex to the sitting surface of the subject. The subject is made to sit on a stool with his/her vertical column as straight as possible, legs hanging freely and head on the Frankfurt Horizontal Plane. The anthropometer is placed at the back and between the two buttocks, taking care that the lumber curve of the subject is not flattened, but concave from behind. The sliding sleeve of the anthropometer is then lowered down towards the middle of the head so that it would touch the vertex. Reading in centimeter and its fraction were recorded.

Bi-acromial diameter: It measures the straight distance between the two acromia in standing position. The measurement is taken from the back of the subject with the Pelvimeter. The subject is asked to keep his/her shoulders straight. If the hand is given to move downward and upward then we find a point where humerus and scapula is joined. The distance between these two points is known as bi- acromial diameter.

Bi-iliac diameter: It measures the straight distance between the two illo-cristalia. (the most lateral points on the iliac crests). The measurement is taken from the back of the subject with the Pelvimeter. While taking bi-iliac diameter, the subject is asked to stand in natural position.

Head circumference: It measures the maximum circumference of the head taken in one horizontal plane, that is, from glabella to opisthocranion to glabella. This measurement was taken with a tape (precision – 1mm).

Mid upper arm circumference: For measurement of mid upper arm circumference, the subject stand erect, with the arms hanging freely at the sides of the trunk and palms towards the thighs. The midpoint between the lateral tip of acromian and most distal point on the olecranon process of the ulna is located and marked. At the marked point, the tape (flexible and non-elastic) is snug to the skin but not compressing soft tissues, the circumference is recorded to the nearest 0.1 cm.

Waist circumference: It measures the circumference of the abdomen at the most lateral contour of the body between the lower margin of ribs and the superior anterior illiac spine. The subject is asked to stand erect and to keep his/her feet close to each other. The measurement is taken with a tape (flexible and non- elastic) at the right angle to the axis of the body when the subject exhaled normally.

Hip circumference: It measures the circumference of the hips at their widest position over the greater trochanters. The subject is asked to keep his/her feet close to each other and stand erect. The measurement is taken with a tape (flexible and non- elastic) at the right angle to the axis of the body when the subject exhaled normally.

Calf circumference: It measures the circumference of the calf muscles where it is most developed. The measurement is taken on a plane perpendicular to the long axis of the calf. Calf circumference is taken while the subject sits on a table with the legs hanging freely or the subject stands with the feet separated about 20 cm and body weight distributed equally on both feet. This measurement is taken on supine position or with the knee flexed at 90 degree in case of children.

Skin fold thickness at biceps: It measures the thickness of a vertical fold on the front of the upper left arm, directly above the centre of the cubital fossa, at the same level marked on the skin for the upper arm circumference. The Holtain skin fold caliper is held in the right hand. A vertical fold of skin and subcutaneous tissue is picked up gently with the left thumb and index finger, approximately 1.0 cm proximal to the arced level, and the tips of the calipers are applied perpendicular to the skin fold at the marked level. Measurements are recorded to the nearest 0.2 mm.

Skin fold thickness at triceps: The triceps skin fold is measured in the mid line of the posterior aspect of the arm, over the triceps muscle, at a level of mid way between the lateral projection of the acromion process at the shoulder and the olecranon process of the ulna. The midpoint is determined as done in mid upper arm circumference.

Skin fold thickness at sub-scapula: For the measurement of sub-scapula skin fold, the subject stands erect with the upper extremities relaxed. It measures the back beneath the inferior angle of the left scapula with the fold either in a vertical line or slightly inclined downward and laterally in the natural cleavage line of the skin.

Skin fold thickness at supra iliac: When the subject stands in an erect posture, this measurement is taken in the mid axillary line immediately superior to the iliac crest. The skin fold is picked up approximately 1 cm above and 2 cm medial to the anterior superior iliac spine.

Skin fold thickness at calf: It measures the skin fold at the level of maximum calf circumference parallel to the long axis of the calf on its medial aspect. The subject is asked to sit with the knee flexed on the side, which is to be measured, and sole of the corresponding foot on the floor. The skin fold is picked up vertically at the level of the maximum calf circumference, which is marked.

Width of humerus: For this measurement, the subject's elbow is bent to the right angle and the width across the outermost points of the lower end of the humerus is taken.

Width of femur: For this measurement, the subject sits on a table with knees bent to the right angle and the width across the outer most parts of the lower end of the femur is measured.

Age: The age of the subject is categorized as given below:

TABLE II. CLASSIFICATION OF AGE

Age1	19-28 years
Age2	29-38 years
Age3	39-48 years
Age4	49-58 years

Body mass index: Body mass index is calculated as body weight in kilogram divided by height in meters squired. It is an indicator of overall obesity. The cut off points recommended for Asia Pacific region [13] was followed:

TABLE III. CLASSIFICATION OF BMI BASED ON ASIA PACIFIC REGION

CED (Chronic Energy Deficiency) Grade III	BMI < 16.0
CED Grade II	BMI 16.0 – 16.9
CED Grade I	BMI 17.0 – 18.4
NORMAL	BMI 18.5 - 22.99
OVERWEIGHT	BMI 23.0 - 24.99
OBESE	BMI > 25.0

IV. METHODOLOGY

Linear discriminant analysis (LDA) and Logistic Regression (LR) are widely used multivariate statistical methods for classification of data having categorical outcome variables. Both of them are appropriate for the development of linear classification models. However, the two methods differ in their basic idea while Logistic Regression (LR) makes no assumptions on the distribution of the explanatory variables, Linear discriminant analysis (LDA) has been developed for normally distributed explanatory variables. Moreover, if any one of the explanatory variables is categorical in nature, no question of checking normality arises. Keeping these points in mind, Logistic Regression (LR) has been used to classify the three communities of Assam based on the information provided by explanatory variables. For the binary valued output the form the regression model is

$$\operatorname{logit}(\mathbb{E}[Y_i \mid \mathbf{X}_i]) = \operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}\cdot\mathbf{X}_i$$

Where P_i = Probability of a person i belonging to Mising / Deori Community (C=1)

1-P_i = Probability of a person i belonging to Chutia Community (C=0)

X is the matrix of the explanatory variables, β is the parameter vector to be stimated.

In our present study, two logistic regression models have been fitted separately—one to discriminate between Deori and Chutia and the other to discriminate between Mising and Chutia. The popular and efficient method of estimating the logistic regression model is the maximum likelihood method. Nowadays, using statistical software packages like SPSS, SAS etc. one can estimates the parameters of this model. However, to use the above software data should be as per the particular software requirements. As the R Programming is command driven statistical package, the package may be utilized as per the requirement of the user.

R programming is a freeware which is object oriented in nature and similar to S-Plus. It has excellent graphical capabilities and supported by large user networks. It has large number of contributed packages which may be downloaded as and when needed. It is basically used for statistical computations. R is also used by the data miners for data analysis. R may be interfaced with C/C++ and Python for increased speed or functionality.

In the 1990s, Ross Ihaka and Robert Gentleman, statisticians at the University of Auckland in New Zealand, had developed the R Programming Language to perform data analysis. R got its name from its developers' initials, although it was also a reference to the most widely used coding language at the time, S. R has a command line interface and the user writes the commands and expects R to execute it.

Recent Studies and polls also proves that the popularity of R increased substantially. [2][5][11].

R Studio is an Integrated Development Environment for R Programming Language. It is an open source and free. R Studio is available in two editions: R Studio Server and R Studio Desktop. R Studio Desktop is a desktop application and runs locally. R Studio Desktop is available for Linux, Mac OS X and Windows. R Studio Server runs on Linux and it can be accessed through browser remotely. R Studio uses Qt for its GUI and is written in C++ Language.

Considering all the above points in mind, the R Programming is used to fit the proposed model. For analysis of stratified categorical data, the Cochran Mantel Haenszel test is used. The test allows the comparison of two groups on a categorical response. When there are three nominal variables, out of them the two variables are of 2x2 contingency table format and the third variable that identifies the repeats.

V. EXPERIMENTS

The two experiments were carried out by using R programming. In the first experiment logistic regression was carried out to find the significance among the anthropometric characteristics of the 3 communities of Assam, India. First, the authors considered the Deori and Chutia communities of Assam. The output of the logistic regression is as follows:

```
    Deviance Residuals:

   Min      1Q   Median     3Q     Max
-2.7121  -0.3556  0.2390  0.5857  2.6311
    Coefficients:
```

TABLE IV. LR RESULTS OF DEORI VERSUS CHUTIA COMMUNITY

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.177670	5.660881	-0.561	0.574567
Age2	-0.367419	0.296367	-1.240	0.215071
Age3	-0.233198	0.313183	-0.745	0.456511
Age4	0.010300	0.420320	0.025	0.980451
Weight	-0.132007	0.039232	-3.365	0.000766 ***
Systolic.BP	0.100352	0.013087	7.668	1.75e-14 ***
Diastolic.BP	0.062551	0.017580	3.558	0.000374 ***
Height	0.122208	0.027227	4.488	7.17e-06 ***
Sit.height	0.073310	0.030248	2.424	0.015368 *
Biacromial diameter	-0.179647	0.048948	-3.67	0 0.000242 ***
Biiliac diameter	-0.020336	0.040413	-0.503	0.614824
Head.cir	-0.304476	0.083361	-3.652	0.000260 ***
MUAC	0.372274	0.089043	4.181	2.90e-05

				***
Waist.cir	-0.122558	0.021081	-5.814	6.11e-09 ***
Calf.cir	0.291251	0.068478	4.253	2.11e-05 ***
Skin.calf	-0.057455	0.039546	-1.453	0.146260
Biceps skinfold	0.199083	0.074960	2.656	0.007911 **
Triceps skinfold	-0.181985	0.057537	-3.163	0.001562 **
Subscapula skinfold	0.218092	0.044720	4.877	1.08e-06 ***
Suprailiac skinfold	0.046265	0.040022	1.156	0.247681
Wdth.humerus	-1.392893	0.229518	-6.069	1.29e-09 ***
Wdth.Femur	-0.009992	0.152274	-0.066	0.947681

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

The output of the logistic regression when considered the Mising and Chutia communities of Assam.

Deviance Residuals:

```
    Min       1Q   Median      3Q      Max
-2.9707  -0.7791   0.3423   0.7810   3.5389
```

Coefficients:

TABLE V. LR RESULTS OF MISING VERSUS CHUTIA COMMUNITY

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.338003	4.295615	1.475	0.140089
Age2	-0.099680	0.206643	-0.482	0.629538
Age3	0.339182	0.229022	1.481	0.138606
Age4	0.575389	0.298972	1.925	0.540285
Weight	-0.008793	0.032861	-0.268	0.789016
Systolic.BP	0.009234	0.007021	1.315	0.188434
Diastolic.BP	-0.011556	0.012501	-0.924	0.355293
Height	0.058925	0.021815	2.701	0.006910 **
Sit.height	-0.084252	0.031854	-2.645	0.008170 **
Biacromial diameter	-0.031969	0.030291	-1.055	0.291235
Biiliac diameter	-0.003929	0.029722	-0.132	0.894840
Head.cir	-0.304744	0.061485	-4.956	7.18e-07 ***
MUAC	-0.104435	0.064561	-1.618	0.105745
Waist.cir	0.004070	0.015706	0.259	0.795539
Calf.cir	0.263649	0.053297	4.947	7.55e-07 ***
Skin.calf	-0.093570	0.024172	-3.871	0.000108 ***
Biceps skinfold	0.044137	0.040627	1.086	0.277300

Triceps skinfold	0.035895	0.041581	0.863	0.387988
Subscapula skinfold	0.082051	0.027760	2.956	0.003120 **
Suprailiac skinfold	-0.082798	0.026020	-3.182	0.001462 **
Wdth.humerus	-1.102172	0.166099	-6.636	3.23e-11 ***
Wdth.Femur	1.411180	0.117651	11.995	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

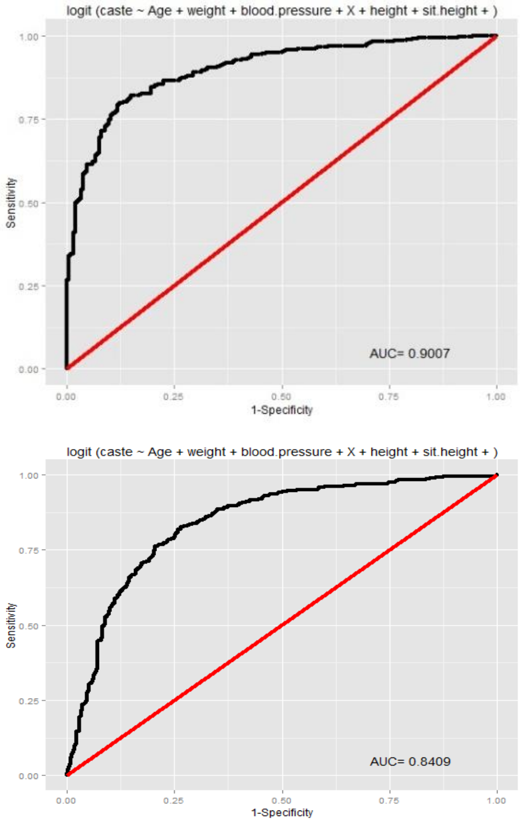


Fig. 1. ROC Curve of Anthropometric Characteristics among Deori and Chutia Community And Mising and Chutia Community respectively

The receiver operating characteristic (ROC) curve is used to measure the performance of classification.The area under the curve as shown figure 1 is 0.9007 for the Deori versus the Chutia communities of Assam. The area under the curve is 0.8409 for the Mising versus the Chutia Communities of Assam.

The second experiment is Cochran-Mantel-Haenszel test. The overweight and obese are considered based on the BMI of the persons. The BMI with the normal range are termed as normal in the following matrix. The row may read as there are 36 persons from Mising communities who are overweight and obese and 113 are in normal category according to the BMI chart as described above. The following is the matrix for the persons whose age is less than or equal to 40.

TABLE VI. MATRIX 1 FOR THE PERSONS OF COMMUNITIESWITH AGE <= 40

	Overweight & Obese	Normal
Mising	36	113
Deori	30	50
Chutia	58	139

Another matrix is for the persons belonging to three different communities of Assam whose age is above 40.

TABLE VII. MATRIX 2 FOR THE PERSONS OF THE COMMUNITIES WITH AGE> 40

	Overweight & Obese	Normal
Mising	39	187
Deori	33	81
Chutia	85	148

The result of the Cochran-Mantel-Haenszel test is as follows:-

CMH statistic = 214.4800, df = 1.0000, p-value = 0.0000, MH Estimate = 2.4808, Pooled Odd

Ratio = 2.4975, Odd Ratio of level 1 = 2.3379, Odd Ratio of level 2 = 2.6000

It is observed that the odds ratio for the first stratum (age below or equal to 40) is 2.3379, the odds ratio for the second stratum (age above 40) is 2.3379, and the aggregate odds ratio that we would get if we pooled the data for both the group is 2.4975. The Mantel-Haenszel odds ratio is estimated to be 2.4808.

VI. DISCUSSION

For classifying the Deori and Chutia communities of Assam, the statistically significant anthropometric characteristics are found to be weight, blood pressure, height, sitting height, bi-acromial diameter, head circumference, mid upper arm circumference, waist circumference, calf circumference, skin fold thickness at biceps, skin fold thickness at triceps, skin fold thickness at sub-scapula and width of humerus. The Table IV reveals that most of the explanatory variables considered in our study can be taken as determinants to discriminate between Chutia and Deori community. It is observed that out of 21 explanatory variables, 14 variables can be used to classify whether a particular person is belonging to Chutia or Deori community. As an illustration, if we consider the parameter weight, it is observed that on average the weight of a Deori community is different from Chutia community. For classifying the Mising and Chutia communities of Assam, the statistically significant anthropometric characteristics are found to be height, sit height, head circumference, calf circumference, skin fold thickness at calf, skin fold thickness at sub-scapula, skin fold thickness at supra iliac, width of humerus and width of femur. Similarly, using logistic regression analysis, to classify between Mising and Chutia Community, out of 21 explanatory variables, 9 variables found to be significant as shown in Table V. The area under the curve for the ROC curve among the communities of Assam also proved it to be good classifiers.

Although the Deori were considered as the priestly section and one of the major sub-division of the Chutiyas [3] but the present study reveals marked difference between them. In this regard, the influence of physical environment factors on the present quantitative traits may not be so significant because these populations are by and large living in a similar ecological condition in upper Assam. Therefore, the difference may be owing to the absence of relatively less gene flow due to strict endogamy as the Deori are well known for their religiosity

In the second experiment, the authors used Cochran-Mantel-Haenszel test to study the association between communities and nutritional status (obese and overweight and normal) stratified by the age. The finding reveals that there is association between them.

REFERENCES

[1] Chobanian, A.V.; Bakris, GL; Black, HR; Cushman, WC; Green, LA; and Izzo, JL (Jr)., *The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure: The JNC-VII Report*, Journal of American Medical Association, 282, 19:2560-2572, 2003.

[2] David Smith, *R Tops Data Mining Software Poll*, Java Developers Journal, May 31, 2012.

[3] Gait, E. A., *A History of Assam.* Calcutta: Thacker, Spink & Co. , 1905.

[4] Genkin Alexander, Lewis David D. and Madigan David, *Large-Scale Bayesian Logistic Regression for Text Categorization*, Technometrics, August 2007, Vol. 49, No. 3.

[5] Karl Rexer, Heather Allen, and Paul Gearan, *Data Miner Survey Summary*, presented at Predictive Analytics World, Oct. 2011.

[6] Kitamura Shingo , Enomoto Minori , Kamei Yuichi , Inada Naoko , Moriwaki Aiko , Kamio Yoko and Mishima Kazuo, *Association between delayed bedtime and sleep-related problems among community dwelling 2-year-old children in Japan*, Journal of Physiological Anthropology (2015) 34:12

[7] Liao J.G. and Chin Khew-Voon, *Logistic regression for disease classification using microarray data: model selection in a large p and small n case*, BIOINFORMATICS, Vol. 23 no. 15 2007, pages 1945–1951

[8] Lobo SW , Chandrashekhar TS and Kumar S, *Cephalic index of Gurung community of Nepal - An anthropometric study,* Kathmandu University Medical Journal (2005) Vol. 3, No. 3, Issue 11, 263-265.

[9] Montgomery M E, White M E, and Martin S W, *A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows*, Can J Vet Res. 1987 Oct; 51(4): 495–498

[10] Pohar Maja , Blas Mateja , and Turk Sandra, *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study*, Metodološki zvezki, Vol. 1, No. 1, 2004, 143-161.

[11] Robert A. Muenchen ,"*The Popularity of Data Analysis Software," 2012*

[12] Sina Maryam , Hoy Wendy and Wang Zhiqiang, *Anthropometric predictors of gestational hypertensive disorders in a remote aboriginal community: a nested case–control study*, BMC Research Notes 2014, 7:122.

[13] World Health Organization, Health Situation in South East Asia Region.1998-2000, WHO Regional Office for South East Asia, New York, 2002.

[14] Zhou Xiaobo , Liu Kuang-Yu and Wong Stephen T.C. , *Cancer classification and prediction using logistic regression with Bayesian gene selection*, Journal of Biomedical Informatics, Volume 37, Issue 4, August 2004, Pages 249–259.

Learners' Attitudes Towards Extended-Blended Learning Experience Based on the S2P Learning Model

Salah Eddine BAHJI
Mohammed V University
Mohammadia School of Engineers-Rabat
High School of Technology-Salé
Rabat, Morocco

Jamila EL ALAMI
Mohammed V University
High School of Technology-Salé
Rabat, Morocco

Youssef LEFDAOUI
Mohammed V University
High School of Technology- Salé
Rabat, Morocco

***Abstract*—Within the Moroccan context, the Higher Education Institutions have realized the importance of the integration of information technologies into the formal learning curriculum. However, the risks of demotivation remain large in tertiary education, even with the support of these new technologies. It is therefore essential to ensure consistently the maintenance of the learners' motivation, which must start from the design phase by adopting real motivational strategies. Blended Learning addresses the issue of the quality of teaching and learning offering then some answers to learners' motivation issue. So, we try to extend of the dimensions of Blended Learning to an "Extended Blended Learning (Ex-BL)", according to the S2P Model Learning designed as an integration model, arguing that knowledge and learning tools are nowadays available everywhere. This integration of educational resources takes into consideration various components: Face-to-face/online learning, Text-based Learning/Game-based Learning/Media-based Learning, Gamification, and Open Educational Resources. This paper investigates the learner perceptions of this instructional design. This includes their perceptions of learning effectiveness and its impact on their motivation during the learning experience. This investigation focused on two main points: the "Observation of learners' behavior", especially during online activities, as a way to gauge the degree of motivation and engagement; and the "Evaluation of the learning experience" through a survey covering the appreciation of the instructional design; the degree of satisfaction; the students' motivation; the online platform; the extension of Ex-BL elements and their impact on learners' motivation.**

***Keywords*—learning model; S2P learning model; blended learning; extended-blended learning; learning experience; learners' motivation; gamification**

I. INTRODUCTION

It is noted a significant movement within the sphere of Higher Education in general, working to meet the needs of the 21st century. In this movement, the Higher Education Institutions have realized the importance of the integration of information technologies into the formal learning curriculum.

Today, existing learning systems are largely focused on the administration of courses and content. Thus, for years, it was recommended to support individual and collective learning processes by using online learning systems [1].

However, the risks of demotivation remain large in tertiary education, even with the support of new technologies. It is therefore essential to ensure consistently the maintenance of the learners' motivation, which must start from the design phase by adopting real motivational strategies [2].

According to Garrison D.R. and Vaughan N.D., Blended Learning addresses the issue of the quality of teaching and learning. This quality is especially visible at the level of motivation and commitment shown by learners during formal learning process [3]. Therefore, this study was conducted to explore the prospects for fostering the learners' motivation and engagement through extended modes of Blended Learning (***Extended Blended Learning***).

Thus, we will focus on designing a prototype of a learning experience base of Extended Blended Learning (Face-to-face/Online; Text-based Learning/Game-based Learning/Media-based Learning, Serious Games/Gamification; Own Content/Open Content) proposed as part of the "*S2P Learning Model*" [4];[5];[6];[7].

To answer our research questions we chose to conduct a survey, with the various target learners in the learning experiment order to evaluate their level of satisfaction (or dissatisfaction) and to observe their degree of motivation and commitment.

II. THEORITICAL BACKGOUND

To approach the underlying theoretical background for our research we will try to outline some key concepts, ranging from definitions to current and perspectives that affect each of the key concepts used in the context of this work. Namely: Motivation; E-Learning & Open Educational Resources; Blended Learning; and Game-based Learning and Gamification.

A. The motivation

Motivation is defined by the Cambridge Academic Content Dictionary as "*the reason to do something*" and the "*willingness to do something, or something that causes such willingness*". Furthermore, Larousse Dictionary defined it as the "*physiological and psychological processes responsible for the initiation, the continuation and the termination of* *behavior*"[8]; while for Maehr M.L. and Meyer H.A., motivation is a "*theoretical construct to explain the initiation, direction, intensity, persistence and quality of behavior, particularly directed behavior toward a goal*" [9].

The theorists distinguish between two dimensions of motivation: "*Intrinsic motivation*" and "*Extrinsic motivation*".

The Intrinsic motivation, which is difficult to master and measure, is intimately linked to the Cognitivism school; while the Extrinsic motivation is supported by behaviourism school, where, according to some stimuli we seek to develop some answers, which can be observed and measured [10].

The Motivation is then seen as the key to effective learning. But due to several factors, keeping learners motivated and engaged during a learning experience is a difficult and deli-cate issue today [11].

B. Technology Enhanced Learning: E-Learning, Open Educational Resources & Blended Learning

In the 21st Century, people need to learn more than ever before. The rapid development of information technologies, especially the Internet has radically improved the manners of teaching and learning [12].

Today, the relationship Technology/Education is becoming more dynamic and interdependent. In this context, education and ICT are the top priorities for the post-2015 development agenda for African countries [13].

Bilinovac D. noted that the concept of e-Learning was not only efficient but more rapid by the emergence of the Internet, arguing that the learning tools are available online, and participants (students and tutors) communicate by e-mail, chat, discussion forums or social networks. Therefore, the concept could be used as a main learning mode or as a combined approach with the class-face [14].

E-Learning results therefore from the association of interactive and multimedia content, distribution media (PC, Internet, Intranet, Extranet), a set of software tools that manage online training and tools for creation of interactive trainings[15]. It is a way that is particularly valuable because it offers more flexibility with lower costs [13].

Furthermore, the use of Open Educational Resources (OER) integrated in the daily life is emerging, even in schools or educational learning landscapes [16]. The OER movement originated from the evolution of Open and Distance Education (ODL) and in a broader context, a culture of open knowledge, Open source, free sharing and peer collaboration, which emerged in the late twentieth century [17].

Also, the evolution of the learning process in teaching was based on the integration of new teaching strategies to improve teaching and increase flexibility. Then, several studies have been conducted to explore learning strategies that exploit the potential of e-Learning, while retaining the advantages of face-to-face teaching, from which the concept of blended learning has emerged. Rooney J.E. said that blended learning has been recognized by the American Society for Training & Development as one of the top ten trends to emerge in the area of knowledge transmission [18].

Indeed, Blended Learning is increasingly offered at Colleges and Universities mainly American [3], with increasing evidence that they can improve student learning [19]. Experience has also shown that well-designed hybrid courses enhance learning of learners and increase knowledge retention [20].

However, the Blended Learning depends largely on technical resources with which the learning experience is designed. These tools must be up to date, reliable and easy in order to have a significant impact. These tools help learners to learn or review key concepts, stay organized, show what they have learned, submit jobs, track achievements and communicate [21].

C. Game-based Learning and Gamification

The popularity of computer games has led to the reflection on their application in education. The games are becoming an integral part of modern society. They are an ideal platform to present new content and new technologies because many people play computer games and accept it as a normal form of entertainment. Contrary to the existing media, games promote interaction, allowing users to participate actively, not passively in receiving information. Therefore, educational games were gradually perceived as very effective tools to improve the teaching/learning activities at higher education [22].

However, after an exploration of gaming activities in education, a new concept was born: Gamification, seen as the use of "*thought*" of the game (*game thinking*) and mechanisms of the game (*game mechanics*) in non-gaming contexts, used to engage users in their activities (problem solving, learning, etc.).

In fact, Gamification is a very recent trend that began to gain ground from 2010 [23]. It mainly involves the use of elements and mechanisms of the game (*game mechanics*) in non-ludic systems (*non-game systems*), in order to improve the experience and engagement of the user.

However, little is known about the efficient design of "*gamified*" systems. The recent introduction of gamified applications to a wide audience of users promises new thinking and rich data sources for the many efforts in Human-Computer Interaction (HMI) that explored mental patterns associated with game, design patterns and motivation dynamics, "funology", social psychology of online communities, or the experiences of users in the game, etc.[24].

Then, the definition presented by Erenli K. remains the most complete, presenting Gamification as follows: *"Gamification is the process of using game mechanics and game thinking in non-gaming contexts to engage users and to solve problems. Gamification leveraged game design, loyalty program design and behavioral economics to create the optimal context for behavior change and successful outcomes"[25].*

III. METHODOLOGY

A. Research objectives

The main goal of this research is to elucidate the factors of design of a motivating learning experience, within the context of tertiary education.

Therefore, this research will focus on developing a learning experience based on Extended-Blended Learning Approach (face-to-face/online; text-based learning/game-based learning/media-based learning; serious games/gamification; own content/open content) proposed according to the framework of the S2P Learning Model [4];[5];[6];[7].

B. Participants

The target population for this study is a community of learners of a Bachelor Degree, within a public institution of higher education (Information Science School, Rabat, Morocco). This population is composed by of 94 students of the 4th year of the cycle "*Informatistes*" (or Specialists in Information Management), engaged in the course "*Knowledge management*". This is a population composed in its majority by female gender, because 78% of students are of the female gender, and 22% are of the male gender. Moreover, the access to the Internet is also a key success factor within our context. Thus, it is observed that 98% of students have a permanent access to the Internet, while only 2% of students do not. Therefore, they can access from the School or Internet cafes.

C. Instruments

To answer our research questions, we have chosen the survey approach in order to evaluate the level of satisfaction of our learners, and to observe their levels of motivation and engagement. We have favored the Investigative technique using data collection instruments like "*Observation Grid*" and "*Questionnaire*".

1) The observation grid

Since the behavior of learners in the classroom was more or less mastered, we focused in our approach, on observing online behavior through the Edmodo Platform (www.edmodo.com).

This observation was conducted according to the following three main areas:

- consultation of resources shared by the teacher.
- participation of students through shared resources and comments.
- gamification elements.

2) The questionnaire

The evaluation of the learning experience was designed and conducted regarding the following aspects: the motivation; the online platform; the game elements & Gamification; and the general appreciation of learners.

The questionnaire was delivered online using Google Forms, due to several factors such as: a facility and the rapidity of forms design; a facility of deployment and implementation of operational forms; a facility of data exploitation (exporting data to other operating or processing tools such as MS Excel and SPSS).

D. Survey context

The learning experience has been evaluated with a class at the "Information Science School", which is the only institution in Morocco that forms the Information Professionals named "*Informatiste*" (1) and "*Informatiste Spécialisé*" (2).

(1) *Informatiste*: Information Professional with a Bachelor Degree.

(2) *Informatiste Spécialisé*: Information Professional with a Master Degree.

This School provides initial training and continuing training in Information Science, including the fields of library and documentation; archives and records management; competitive intelligence and business intelligence; information management; information and knowledge systems and related fields.

IV. DESIGN OF A LEARNING EXPERIENCE ACORDING TO THE S2P LEARNING MODEL: APPLICATION FOR AN EXTENDED-BLENDED LEARNING

The approach for the fight against demotivation of learners is a "top-down" approach based on three levels: a ***Macro level***, focusing on the definition of a Conceptual framework, namely the S2P Learning Model; a ***Meso level***, putting emphasis on the design of the Learning Experience; and a ***Micro level***, putting emphasis on content conception and learning activities [6].

To ensure a proper understanding of any formal learning approach, it was appropriate to draw a logical framework for the definition and implementation of any educational initiative. This framework was designed through the definition of three complementary dimensions: a strategic dimension of reference (***Formal Learning Strategy***), a technical dimension of support (***Formal Learning Platform***), and procedural dimension of knowledge acquisition (***Formal Learning Process***). These three dimensions interact into a relationship framework and dynamic interdependence: "***Definition***" – "***Support***" – "***Adjustment***"[4];[26].

A. The Formal Learning Strategy

The domain – Field: Our experiment was conducted as part of the course entitled "Knowledge Management" given to the final year (4th year) of the Bachelor degree.

The Level: The desired level is a level midway between Initiation and Development, because students will be exposed to new concepts and new theories related to Knowledge Management. In addition, they will have to improve the use of information management tools in the sphere of the discipline.

The Content: The course content is mostly theoretical, for the internalization and appropriation of concepts. But it is administered in various forms.

The Pedagogical objectives: This course aims to introduce and deepen the students' knowledge about Knowledge Management, which is an extension of the central discipline of their profile, namely: Information Management.

To do this, intermediate objectives are defined, such as: a) understanding the concept of "*Knowledge*" ; b) understanding the *Knowledge Management* in the organization and its methods of approaches ; c) understanding the impacts of the integration of this mode of management on new ways of business organization.

The Pedagogical scenario: The Pedagogical scenario designed for the animation of the course is conducted on six units. Each unit consists on various sections according to the traced educational goals “Fig. 2”.

B. *The Formal Learning Platform*

The fundamental principle adopted in the design of the learning platform is the principle of Integration. Where several components have been integrated such as: course lectures; presentations made by learners; case study analysis; online assignments. Supported by an online platform “*Edmodo*” (used for sharing course materials; sharing web resources: videos, documents, figures and images; sharing comments; integration of online serious games; assigning work); making online works; assessment and evaluation; archiving of all course resources; insertion of Gamification elements (badges assignment; awarding points; rewards).

The simple use of Blended Learning to overcome the problem of classroom learners’ demotivation may be insufficient. Therefore, it would be appropriate to expand the instructional design. An initiative particularly motivated by the development of ICT, promoting an easy access to sources of knowledge. Thus, we propose the “*Extended-Blended Learning*” that will take its full meaning by incorporating the following layers in the “Fig. 1”.

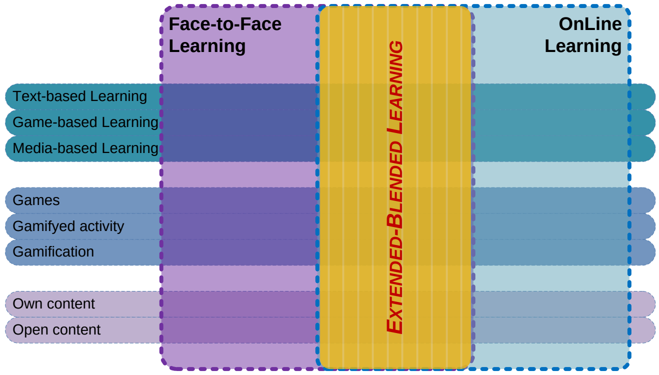


Fig. 1. Components of the “Extended-Blended Learning”

The Extended-Blended Learning exceeds Blended Learning and is characterized by the combination of several components on several levels (in addition to face-to-face and online), including: the pedagogy; the media; the content; the type of contact; the activity; the assessment; etc.

Why focusing on Game mechanics more than games? In the game, the player can win as he can lose; it can be a psychological barrier. While in an educational environment, we do not seek the learner “loses”, but mostly we try to take advantage of the positive aspects of Game Mechanics to raise motivation and commitment. As result, make the maximum that the learner finds the environment for success (“Win”).

The gamification principles can be used in online and face-to-face activities, rewarding any relevant participation and any deserving effort from students.

Therefore, the choice of Game Mechanics in the S2P Learning Model must be in correlation with the learning process we want to initiate within the learner.

C. *The Formal Learning Process*

The overall style of the course focuses on the development of four learning processes (at varying degrees), including: the process of Internalization; the process of Reflection; the process of Decision-making; and the process Socialization.

During the course, the ultimate goal lay in the internalization of concepts, approaches, methods and tools dedicated to KM. Therefore, the knowledge internalization process was the focal point of our approach.

The reflection process was encouraged in the context of analytical work, case studies, preparation of oral presentations.

The process of decision-making was present during serious games, integrated in the course, as well as at intermediate evaluation quizzes deployed at each course unit.

The socialization process was present at the collaborative work assigned to students during the analytical work, case studies and oral presentations.

V. Results and Discussion

The evaluation the learning experience was driven by the combination of two complementary components, namely: the observation of learner behavior and the evaluation via a survey.

A. *Observation of learners’ behavior*

The elements of observation of the behavior cover learners’ online activities as a way to estimate the degree of motivation and engagement during the learning experience. It covers:

- the consultation rates of resources shared by the teacher.
- the learners’ participation through the sharing of resources and comments.
- the components of Gamification.

1) *Consultation rates of resources shared by the teacher*

During this learning experience, the teacher shared with students 18 different resources in terms of nature and in terms of goals. The Edmodo platform offers statistics views to monitor the status of consultation of posted documents. The teacher's shared resources have recorded 1752 consultations, with an overall average of 97.33 consultations per document, as seen in “Table 1”.

The consultation rate of these resources varies from one resource to another, depending on the course units and according to the assignment type.

The highest consultation rates concerned resources of type “Assignment” (work, duty, case studies), when resources that do not have an obligatory form recorded lower rates.

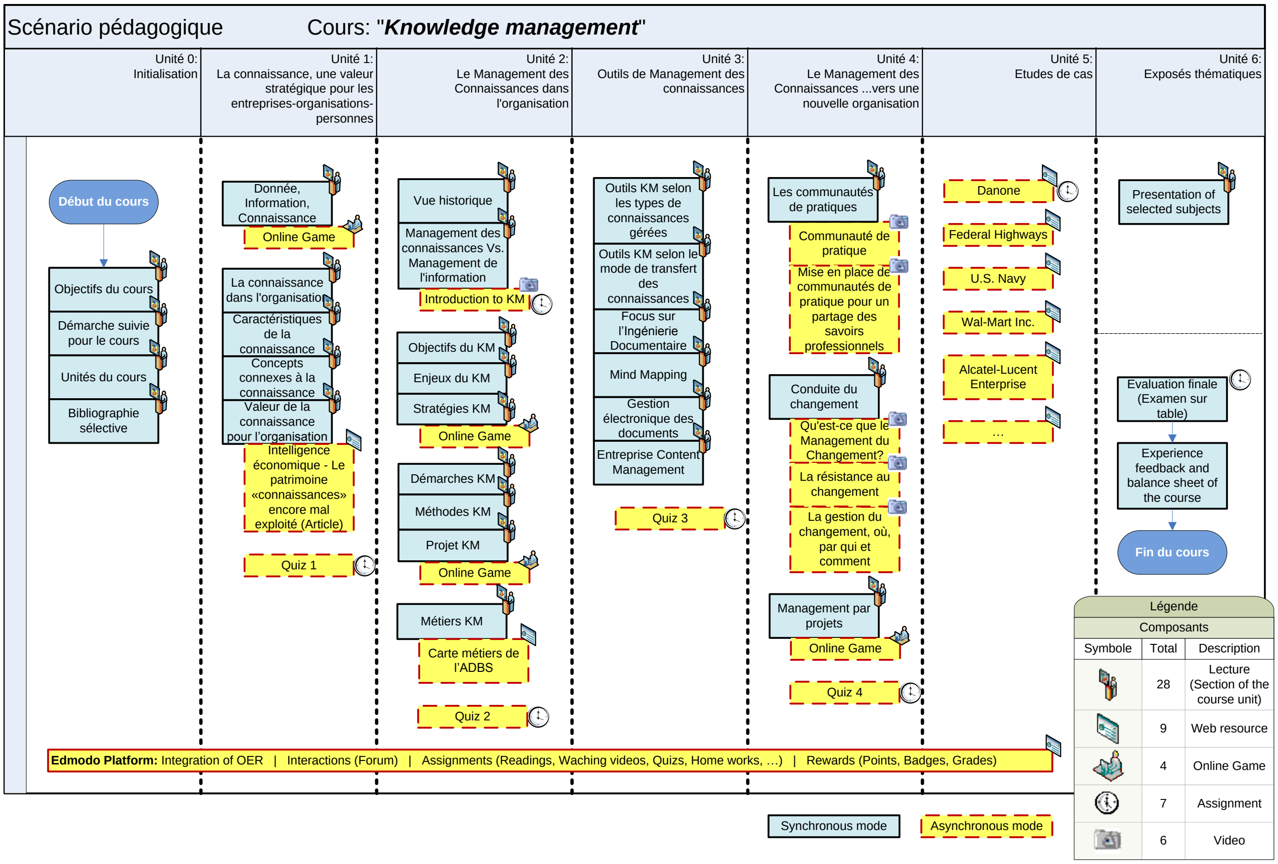

Scénario pédagogique
Cours: "Knowledge management"
Unité 0: Initialisation
Unité 1: La connaissance, une valeur stratégique pour les entreprises-organisations-personnes
Unité 2: Le Management des Connaissances dans l'organisation
Unité 3: Outils de Management des connaissances
Unité 4: Le Management des Connaissances ...vers une nouvelle organisation
Unité 5: Etudes de cas
Unité 6: Exposés thématiques
Début du cours
Objectifs du cours
Démarche suivie pour le cours
Unités du cours
Bibliographie sélective
Donnée, Information, Connaissance
Online Game
La connaissance dans l'organisation
Caractéristiques de la connaissance
Concepts connexes à la connaissance
Valeur de la connaissance pour l'organisation
Intelligence économique - Le patrimoine «connaissances» encore mal exploité (Article)
Quiz 1
Vue historique
Management des connaissances Vs. Management de l'information
Introduction to KM
Objectifs du KM
Enjeux du KM
Stratégies KM
Online Game
Démarches KM
Méthodes KM
Projet KM
Online Game
Métiers KM
Carte métiers de l'ADBS
Quiz 2
Outils KM selon les types de connaissances gérées
Outils KM selon le mode de transfert des connaissances
Focus sur l'Ingénierie Documentaire
Mind Mapping
Gestion électronique des documents
Entreprise Content Management
Quiz 3
Les communautés de pratiques
Communauté de pratique
Mise en place de communautés de pratique pour un partage des savoirs professionnels
Conduite du changement
Qu'est-ce que le Management du Changement?
La résistance au changement
La gestion du changement, où, par qui et comment
Management par projets
Online Game
Quiz 4
Danone
Federal Highways
U.S. Navy
Wal-Mart Inc.
Alcatel-Lucent Enterprise
...
Presentation of selected subjects
Evaluation finale (Examen sur table)
Experience feedback and balance sheet of the course
Fin du cours
Edmodo Platform: Integration of OER | Interactions (Forum) | Assignments (Readings, Waching videos, Quizs, Home works, ...) | Rewards (Points, Badges, Grades)
Synchronous mode
Asynchronous mode
Légende
Composants
Symbole
Total
Description
28
Lecture (Section of the course unit)
9
Web resource
4
Online Game
7
Assignment
6
Video

Fig. 2. Components of the pedagogical scenario

TABLE I. CONSULTATION OF RESOURCES SHARED BY THE TEACHER

Types of posted resources	Number of posted resources	Number of consultations	Average of consultations
Slides	5	544	108.8
Online games	2	146	73
Documents	5	446	89.2
Quizzes	3	282	94
Videos	3	334	111.33
	18	***1752***	***97.33***

The most-viewed items are course materials with a total of 544 consultations, followed by documents with 446 consultations, videos with 334 consultations, and finally online games with a total of 146 consultations.

In terms of averages, the videos recorded the highest average by 111.33 consultations per video, followed by course material with an average of 108.80 consultations, finally the online games, with an average of 73 consultations per game.

2) Learners' participation through the sharing of resources and comments

Learners' participation through the sharing of resources and comments will be estimated using the participation rate and the type of participation. The learners' participation rate at the online platform is significant, in the order of 75.5% of students (active participation by sharing comments and documents). While nearly a quarter of the students showed a passive presence at the online platform (observers).

Among the 75.5% of students who have shared items at the online platform, it is observed that 52.1% of them have recorded between one to five shares each. 22.5% had between six to ten shares for each student of them, while 25.4% of students have shared more than ten resources each at the online platform.

The learners' contribution at the online platform totaled 827 participations. 67.2% in the form of electronic resources (documents and others), and 32.8% as comments. With a global average of 8.8 participations per student (a ventilated average of 5.9 documents shared by student, and an average of 2.9 comments by students).

3) Components of Gamification

To reward the student's participation during learning activities, whether in class or online, the principle of Gamification activities was adopted by allocating badges. These badges are transformed into points as Rewards.

During this experiment, we registered the distribution of a total of 446 badges for all categories. In fact, it has been observed a positive involvement for the adopted approach, since 65% of students obtained at least one badge.

Thus, we can see that 42 students scored between 1 and 5 badges each (a rate of 45%). 9 students achieved between 6 and 10 badges each (a rate of 9%). In addition, 11 students received more than 10 badges each (at the rate of 11%), while 35% did not get any badge.

The most collected badge is "*Participant*", with a total of 381 distributed badges (a rate of 85%). With a maximum of 37 badges for a student, (an average of 4.05 badges per student and a standard deviation about 7.8). Followed by the type "*Perfect Attendance*" with total of 29 badges distributed (a rate of 7%). "*Homework Helper*" comes in third place with a rate of 3.4% of distributed badges. In the fourth place, "*Hard Worker*" with a rate of about 2% of the distributed badges. In the fifth position, we find "*Good Citizen*" with a rate of 1.3%. In the last positions "*Student of the Month*" and "*Star Performer*" with rates around 1% each, as presented in "Table 2".

In fact, among 63 students (those who have collected badges), more than a half (57.1%) has received at least two different types of badges. While about 43% of the students have collected only one type of badges.

TABLE II. TYPES OF DISTRIBUTED BADGES

Badges	**Total distributed**	**Percentage**	**Average**	**Standard deviation**
Good Citizen	6	1,3%	0,06	0,29
Hard Worker	8	1,8%	0,09	0,28
Homework Helper	15	3,4%	0,16	0,37
Participant	381	85,4%	4,05	7,84
Perfect Attendance	29	6,5%	0,31	0,53
Star Performer	3	0,7%	0,03	0,18
Student of the Month	4	0,9%	0,04	0,20
	446	***100%***	***4,74***	***8,22***

Among those who got two types of badges and more, 35% received two different types of badges. 17% received three different types of badges and 5% received between four to five different types of badges. Thus, an overall average of 1.9 different types of badges collected per learner was recorded.

B. *Evaluation of the learning experience*

The evaluation of the learning experience has considered the following: the appreciation of instructional design by students; the appreciation of the degree of students' satisfaction; the appreciation of students' motivation; the appreciation of the online platform; the perception of the extension of Ex-BL elements and their impact on learners' motivation.

1) Appreciation of the instructional design applied to the courses

On the question related to the impact of the instructional design applied to the Course on the encouragement of the learning experience, it should be noted that the vast majority of students is convinced.

The components of this instructional design are appreciated positively since the majority is perceived as "*Very interesting*" and "*Interesting*", as presented in "Fig. 3".

This evaluation has considered the following components: Units of the course; Course content; Blended learning mode; online platform (Edmodo.com); Game elements embedded; Learning through play.

We record a positive overall appreciation of the Units of the course, because 83% of students find these units "*Very interesting*" and "*Interesting*", while 15% of students are the "*Interesting enough*".

The Course content recorded the same appreciation as 85% of students find it "*Interesting*" to "*Very interesting*", while only 15% find the "*Interesting enough*".

Students' perception of the Blended learning approach was positive, the fact that 94% of students consider it "*Very interesting*" and "*Interesting*", while only 6% of students who consider it "*Less interesting*".

For the Online platform used in the course, a large majority of students (92%) perceives it "*Very interesting*" and "*Interesting*", while 4% of students judge it "*Interesting enough*". Only 4% of students have a negative perception of the platform, since 2% sees it "*Less interesting*", and 2% sees it "*Not at all interesting*".

The integration of Game mechanics at the learning experience has also attracted the interest of students. Indeed, 76% of students felt "*Very interesting*" to "*Interesting*". 9% of the students find then "*Interesting enough*", while 7% of students find the "*Less interesting*" and 2% found them "*Not at all interesting*".

In the same line, Learning through play attracted the same interest, since 76% of students consider it "*Very interesting*" to "*Interesting*". 9% of students find it "*Interesting enough*", while 8% of students find it "*Less interesting*".

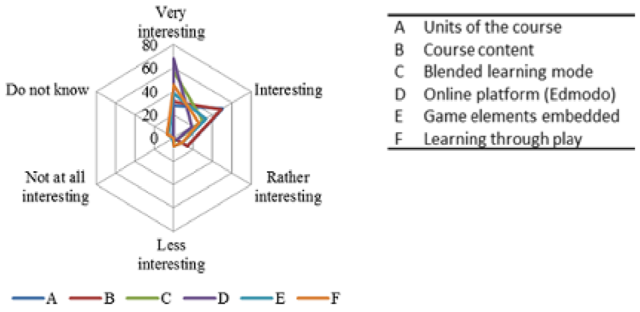


Fig. 3. Perception of the components of instructional design adopted in the course

2) Students' motivation

During the learning experience, it was noted a significant level of motivation as seen in "Fig. 4", because 96% of students consider being motivated: 61% "*Very motivated*" and 35% "*Motivated enough*". Only 2% of students consider being "*Less motivated*".

Several factors (intrinsic and extrinsic) contributed to maintain learners' motivation. Therefore, the majority of students' appreciations regarding the importance of intrinsic motivation factors were concentrated around "*Extremely important*", "*Very important*" and "*Important*":

- *The course will facilitate employability*: 83% of students consider it "*Important*" to "*Extremely important*", and 15% "*Important enough*".
- *The course gives a new dimension to the specialty (Information Management)*: 96% of students consider it "*Important*" to "*Extremely important*", while only 4% of the students consider it "*Important enough*".

In parallel of intrinsic motivation factors, students were approached in relation to extrinsic motivators developed in this learning experience.

The overall view of these factors shows a positive perception since the majority assessments were concentrated around "*Extremely important*", "*Very important*" and "*Important*":

- *The Blended mode*: was appreciated as "*Important*" to "*Extremely important*" by 95% of students.
- *The Teacher's style*: was considered as "*Important*" to "*Extremely important*" by 94% of students.
- *The Teacher's personality*: was appreciated as "*Important*" to "*Extremely important*" by 95% of students.
- *Learning through play*: has attracted the interest of 84% of students; "*Important enough*" to "*Extremely important*".
- *Game mechanics*: has been "*Important*" to "*Extremely important*" to 82% of students, and "*Important enough*" to 6%.

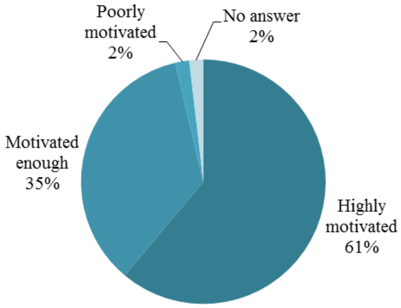


Fig. 4. Learners' motivation level during the learning experience

- *The Exchanges between students*: in class 97%, online 86%, "*Important*" to "*Extremely important*".
- *The Exchange with the teacher*: in the classroom 99%, Online 92%, "*Important*" to "*Extremely important*".
- *The Integration of Open Educational Resources*: 93% of students found it "*Important*" to "*Extremely important*".

C. Discussion

The objective of this study is to provide more empirical researches on the effects of the extended blended learning; and to measure the effectiveness of the extended mode on the engagement and motivation of learners.

Previous research work on the impact of Blended Learning on performance, motivation and engagement revealed that learners who participate in both synchronous, as asynchronous modes are more engaged and also demonstrate a significant improvement in skills [27].

Two main effects are highlighted in this research: the effect related to the extended design (*Extended Blended Learning*), and the principle of integration of educational resources. Both in terms of the observations and the evaluation of the learning experience, the results obtained in this research are encouraging.

1) Instructional design based on Ex-BL

The components of the adopted instructional design are appreciated positively. This appreciation has considered components as varied as the units of the course; the course content; the blended learning mode; the online platform used (Edmodo), Gamification and Game mechanics; and serious games.

2) Learners' motivation

Without prior knowledge of the subject matter and without prior knowledge of the online platform Edmodo, learners have demonstrated sustained motivation throughout this learning experience.

This motivation was further encouraged by the adoption of the Ex-BL as instructional design approach, the fact that learners have a positive perception for the extension of Ex-BL in formal learning and its impact on learners' motivation.

Moreover, the motivation of learners was apprehended at both the observation of behavior and through the evaluation of the learning experience.

Indeed, the observations made from this experiment indicate a strong involvement of learners in the learning process. A positive implication has been recorded through various indicators, including:

- a significant rate of student participation.
- a rich and varied electronic resource sharing.
- high consultation rates of shared resources.
- rich participation by issuing comments on shared resources.

This involvement was particularly encouraged by the Game mechanics & Gamification embedded in the learning process, ensuring a reward to any significant and beneficial learner's participation. Then, learners appreciated positively the principle of "rewarding the effort" of sharing and participation.

Moreover, serious games are a force for technology-enhanced learning [28], the fact that the vast majority of learners appreciated and encouraged at the same time extending their use in formal learning curriculum.

However, the design of educational games is not a simple task and there are no solutions for all uses [29]. In this sense, the Gamification can be a valuable complement.

3) S2P Learning Model: a model for the Extended Blended Learning

In general, the results obtained in this research confirm at a significant extent the assumptions made regarding the use of Extended Blended Learning through the S2P Learning Model as canalization vector of learners' motivation and engagement.

Indeed, this model allowed the design of an extended blended learning experience, combining resources as varied as proprietary content, open content, as well as various mediums such as face-to-face, Online, text, video game and Gamification.

VI. Conclusions

In the context of this study, we showed that the S2P Learning Model as a design framework of formal learning experiences based on extensive mixed approach (Extended Blended Learning) presents indicators fostering motivation and engagement of learners. The potential of Extended-Blended Learning remains enormous and still largely untapped. Therefore, we still need a rich conceptual framework integrating both the Gamification, Serious Game, media resources and OER in general, as part of a formal learning initiative, including the S2P Learning Model.

Indeed, the potential of the Text-based Learning combined with the Game-based Learning, and Media-based Learning is enormous. It remains to deepen the elements of an effective combination, because we need to vary teaching tools for effective learning that brings the learner resistance against demotivation and boredom.

In addition, the context plays an important role in the learning process and, therefore, we need to assess the applicability of the model S2P learning in different contexts (personal, organizational context, academic context, etc.).

In any case, the main issue is the ability to find a balance between the formal and informal dimensions of learning. In a sense that the learner can get the most out of each dimension, encouraging individual initiative by developing autonomy and “self-care” instead of develop total dependence.

Therefore, the personal dimension can feed the formal dimension by analysing the learner's behavior outside the formal system and try to integrate individual significant elements in the formal program.

In this sense, the role of the teacher is to watch over the individual practice of learning to try to capture the significant signals from the individual level, to capitalize on the formal dimension shared between learners without weighing the formal program and educational curriculum. In this case, we must go further deeper, taking into account the behavioral and cognitive studies to enrich and expand the model.

References

[1] M.M. Gasland, “Game Mechanics based E-Learning: A case study”. Master thesis Master of Science in Computer Science, Norwegian University of Science and Technology, June 2011.

[2] J. Mignon and J.L. Closset, “Maintien et stratégies de renforcement de la motivation des étudiants dans l’enseignement à distance”. Actes du 21ème congrès de l’Association Internationale de Pédagogie Universitaire (AIPU), du 3 au 7 mai 2004, à Marrakech-Maroc.

[3] D.R. Garrison and N.D. Vaughan, “Blended Learning in Higher Education: Framework, Principles, and Guidelines”. San Francisco: Jossey-Bass, 2008.

[4] S.E. Bahji, Y. Lefdaoui and J. El Alami, “The Learning Model S2P as a Conceptual Framework for Understanding the Serious Game”. Proceedings of the 14th IASTED International Conference “Computers and Advanced Technology in Education (CATE 2011)”.11-13 July 2011, Cambridge - United Kingdom.

[5] S.E. Bahji, Y. Lefdaoui and J. El Alami, “S2P Learning Model for combining Game-Based Learning and Text-Based Learning”. Proceeding of the 5th Guide International Conference 2011 “E-learning innovative models for the integration of education, technology and research”.18-19 November 2011, Rome - Italy.

[6] S.E. Bahji, Y. Lefdaoui and J. El Alami, “Enhancing Motivation and Engagement: A Top-Down Ap-proach for the Design of a Learning Experience According to the S2P-LM”. In: International Journal Of Emerging Technologies In Learning (IJET), Volume 8, Issue 6, December 3, 2013. pp. 35-41.

[7] S.E. Bahji, Y. Lefdaoui and J. El Alami, “The S2P learning model: For the combination of the formal and the personal dimensions of learning”. In: Journal of Mobile Multimedia (JMM), Volume 9, Issue 3-4, March 2014. pp. 242-252.

[8] H. Mouaheb, A. Fahli, M. Moussetad and S. Eljamalic, “The serious game: what educational benefits?”. In: Procedia - Social and Behavioral Sciences. Issue 46, 2012. pp 5502 – 5508.

[9] M.L. Maehr and H.A. Meyer, “Understanding motivation and schooling: Where we’ve been, where we are, and where we need to go”. In: Educational Psychology Review. Volume 9, Issue 4, 1997. pp: 371–409.

[10] F.H. Muller and J. Louw, “Learning environment, motivation and interest: perspectives on self-determination theory”. In: South African journal of psychology, 34 (2), 2004. pp. 169-190.

[11] J.E. Brophy, “Motivating Students to Learn”. New York: Routledge, 3rd Edition, 2010.

[12] I. Ismail, R.M. Idris, H. Baharum, M. Rosli and A. Abu Ziden, “The Learners’ Attitudes towards Using Different Learning Methods in E-Learning Portal Environment”. In: International Journal of Emerging Technologies in Learning - iJET, Volume 6, Issue 3, September 2011. pp. 49-52.

[13] S. Isaacs, D. Hollow, B. Akoh and T. Harper-Merrett, “eLearning Africa Report 2013”. Isaacs S (ed.), ICWE: Germany, 2013.

[14] D. Bilinovac, “ELearning as a Tool for Knowledge Management”. Proceedings of the 14th International Research/Expert Conference: Trends in the Development of Machinery and Associated Technology TMT 2010, Mediterranean Cruise, 11 – 18 September 2010. pp. 381-384.

[15] M. Favier, M. Kalika and J. Trahand, “E-Learning/E-Formation: implications pour les organisations”. In: Systèmes d’Information et Management, Vol. 9, N°4, Décembre 2004.

[16] E.S.I. Ossiannilsson and A.M. Creelman, “OER, Resources for learning – Experiences from an OER Project in Sweden”. In: European Journal of Open and Distance Learning. Issue 1, 2012.

[17] Commonwealth of learning, “COL’s policy on open educational resources”. URL:www.col.org/progServ/policy/Pages/oer.aspx

[18] J.E. Rooney, “Blending Learning Opportunities to Enhance Educational Programming and Meetings”. In: Association Management, 55(5), 2003. pp. 26–32.

[19] B. Means, Y. Toyama, R. Murphy, M. Bakia and K. Jones, “Evaluation of Evidence-based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies”. Washington, D.C.: U.S. Department of Education, 2009.

[20] K.E. Amaral and J.D. Shank, “Enhancing Student Learning and Retention with Blended Learning Class Guides”, 2010. URL:www.educause.edu/ero/article/enhancing-student-learning-and-retention-blended-learning-class-guides

[21] Ontario Ministry of Education, Defining blended learning, 2014. URL:www.edu.gov.on.ca/elearning/blend.html

[22] G. Zichermann and C. Cunningham, “Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps”. O'Reilly Media, Inc., First Edition, 2011.

[23] J.J. Lee and J. Hammer, “Gamification in Education: What, How, Why Bother?”. In: Academic Exchange Quarterly. Volume 15, Issue 2, 2011.

[24] S. Deterding, M. Sicart, L. Nacke, K. O'Hara and D. Dixon, “Gamification. Using Game-Design Elements in Non-Gaming Contexts”. Proceedings of the 2011 annual conference on Human factors in computing systems (CHIEA'11).ACM, New York, NY, USA, 2425-2428.

[25] K. Erenli, “The Impact of Gamification: Recommending Education Scenarios”. In: International Journal of Emerging Technologies in Learning (iJET), Volume 8, Issue 1, Special Issue 1: “ICL2012”, January 2013.

[26] S.E. Bahji, Y. Lefdaoui and J. El Alami, “The Learning Model S2P: a formal and a personal dimension”. Proceeding of the 4th International Conference on Next Generation Networks & Services (NGNS'12). 2-4 December 2012, Algarve-Portugal.

[27] R.Y. Jee and G. O’Connor, “Evaluating the Impact of Blended Learning on Performance and Engagement of Second Language Learners”. In: International Journal of Advanced Corporate Learning (iJAC). Volume 7, Issue 3, 2014. pp. 12-16.

[28] P. Moreno-Ger, D. Burgos and J. Torrente, “Digital Games in eLearning Environments: Current uses and emerging trends”. In: Simulation & Gaming. Volume 40, Issue 5, 2009. pp. 669-687.

[29] P. Moreno-Ger, D. Burgos, I. Martínez-Ortiz, J.L. Sierra, and B. Fernández-Manjón, “Educational Game Design for Online Education”. In: Computers in Human Behavior, 24 (2008). pp. 2530–2540.

Hybrid Algorithm for the Optimization of Training Convolutional Neural Network

Hayder M. Albeahdili
Dep. of Electrical and computer Engineering
University of Missouri, Columbia
Columbia, Missouri, 65211, USA

Tony Han
Dep. of Electrical and computer Engineering
University of Missouri, Columbia
Columbia, Missouri, 65211, USA

Naz E. Islam
Dep. of Electrical and computer Engineering
University of Missouri, Columbia
Columbia, Missouri, 65211, USA

Abstract—**The training optimization processes and efficient fast classification are vital elements in the development of a convolution neural network (CNN). Although stochastic gradient descend (SGD) is a Prevalence algorithm used by many researchers for the optimization of training CNNs, it has vast limitations. In this paper, it is endeavor to diminish and tackle drawbacks inherited from SGD by proposing an alternate algorithm for CNN training optimization. A hybrid of genetic algorithm (GA) and particle swarm optimization (PSO) is deployed in this work. In addition to SGD, PSO and genetic algorithm (PSO-GA) are also incorporated as a combined and efficient mechanism in achieving non trivial solutions. The proposed unified method achieves state-of-the-art classification results on the different challenge benchmark datasets such as MNIST, CIFAR-10, and SVHN. Experimental results showed that the results outperform and achieve superior results to most contemporary approaches.**

Keywords—Convolutional Neural Network; Particle Swarm optimization; Image Classification

I. INTRODUCTION

The Convolutional Neural Network (CNN) algorithm has been widely applied in many applications, including face recognition [1, 2], image classification and recognition [3-6] and object detection [7]. In supervised learning, Back Propagation (BP) algorithm is the prevalence and constituent method used for CNN training and parameters tuning. All researchers used it in CNN training in all their implementations.

However, there are a number of disadvantages of using the back propagation algorithm alone. For example, BP algorithm deterministically occurs in local optima, making it hard to get global optima, especially if a large search space is required for optimal solution. The algorithm is also slow and hardly benefits of using modern machines such as Graphics Processing Unit (GPUs), which runs hundreds to thousands of threads simultaneously. The complex computational equations emerging in the algorithm demand hard and complicated series of steps to find derivative equations for updating weight parameters. Finally, the cardinality of back propagation algorithm recruits intermediate variables to preserve the validity of data. Means, the implication of BP requires keeping forward and backward essential parameters used for updating equations.

To tackle limitations mentioned accompanied with BP algorithm, in this paper an alternate algorithm is proposed for CNN training. In particular, the Particle Swarm Optimization (PSO) algorithm is introduced for training; and it is combined with the Stochastic Gradient Descent (SGD) to achieve better results. The computational algorithm proposed delves to avoid occurring in local optimum, is fully parallel, and induces simple equations for CNN training. It is completely adaptable because it does not require any changes in CNN structure when some network layers are added or eliminated. The PSO equations used for training weights are completely parallelize as described in (1) and (2) and shown in fig. 2. This suggests that the weights of any layer can be updated without the need for backward phase as in SGD, thus GPUs can be completely utilized using this implementation. The proposed method also improves training by overcoming premature saturation and sluggishness inspired by SGD.

The reset of the paper consist of the following: in section II, related works are introduced. In section III a brief introduction of introducing PSO is presented. Then in section IV, the proposed approach is introduced in details. Then in section V the model architecture of CNN is illustrated. In both second VI and VII, challenge benchmark used for model evaluation and conclusion are depicted respectively.

II. RELATED WORK

Recently there are vast number of research have been proposed for image recognition using different methods and several proposed novel methods are proposed. Generally image recognition can be obtained using different approaches such as Pedro F. Felzenszwalb et al. [8] proposed a method for image recognition using Deformable Part Models (DPM). In addition further works are devoted using different strategies of using DPM as demonstrated in [9, 10, 11]. Varity of other methods are used for image classification such as SVM [12, 13, 14,15], boosting [16], spatial pyramid matching [17]; however, on the other hand the most dominant recent works achieved using Convolutional Neural Network (CNN). The last is used widely variety of applications such as image recognition [31, 18, 22, 19, 20], object detection [20, 21, 23, 24], scene labeling [25], segmentation [26, 27], and variety of other tasks [28, 29, 30]. All the mentioned above works use Stochastic Gradient Descent (SGD). However, in this work, this algorithm is replaced by PSO. In addition, hybrid training algorithm of both PSO and SGD is used.

III. PARTICLE SWARM OPTIMIZATION (PSO)

PSO is an evolutionary stochastic optimization computational algorithm introduced by Eberhart and Kennedy [32,33,34]. Particles are randomly initialized, and periodically updated to introduce a new sophisticated population with new fitness. Each particle updates its new position contingent on its history and the best particle history. Thus the particle movement exploits on two values. The first value is the local best, which characterizes the best value so far for the particle itself, and the second value is the global best, which denotes the best value achieved so far by any particle within the swarm. At each time step, particles traverses toward its new best position by altering another parameter termed velocity. The following notions are the formulas used to tune CNN parameters.

$$v_k^{t+1} = w^t v_k^t + c_1 r_1(\mathrm{lbest}_k - x_k^t) + c_2 r_2(\mathrm{gbest} - x_k^t) \qquad (1)$$

$$x_k^{t+1} = x_k^t + v_k^{t+1} \qquad (2)$$

where, v_k^t and x_k^t denote the velocity and position of the particle k at moment t , respectively; c1, and c2 are accelerating factors, r_1, and r_2 are random numbers between [0,1], lbest is the best position for the particle k, gbest is the best particle in the whole swarm. It is obvious that PSO notions are unpretentious and have very rare parameters to be adjusted.

The PSO has remarkable convergence in the initial stages, but it quickly traps to local optimum. In addition, PSO has difficultly incapacitating to avoid local optimum if the search space encompasses only optimal solution [35]. The PSO predominantly experiences premature convergence and searches in region adjacent to global minimum as training progresses chronologically [36, 37] causing PSO permanently trapped in local optimum region. Therefore PSO is amalgamated by Genetic Algorithm (GA), which is an evolutionary algorithm widely used in solving problems in various fields [38-41]. It defines an initial generation that searches in domain space of the problem and generates a new population based mechanisms of reproduction, crossover, and mutation, which is frequently applied to produce new offspring. Usually, new descendants have higher quality and better fitness than ancestors. The GA induces enhancing PSO by merging particles in a bright approach to produce new generations. Combining GA and PSO crucially leverages the proposed hybrid training method by sharing information among particles, increasing the diversity of search space, countenancing the training vital through computation steps, and finally averting PSO to occur in local optimum. To sustain a smooth transition for the hybrid training along computation steps, Genetic Algorithm is applied to PSO whenever there are one of the following factors; i) premature convergence, ii) no progress in the fitness function, or iii) Error changes remains steady from two to three consecutive steps.

IV. PROPOSED OPTIMIZATION ALGORITHM

Since SGD has slow convergence and it cannot be fully parallel to take advantages of GPUs, in this paper, a robust hybrid training algorithm is proposed for CNN training. The algorithm is combined both PSO and SGD, and it is called PSO-SGD, which is a highly parallel method. In this approach, it is expected that the unified PSO and SGD algorithm can crucially achieves superior results and surpass previous methods because of still preserving gains of using SGD and the PSO is recruited as revival constituent. For instance, instead of running one particle, which characterizes the whole CNN parameters, plurality of particles is used and scattered over the scope of search space. Also all particles collaborate with each other using delicate method elucidated in next sections. The proposed training algorithm is divided into dual phase. In the first phase, the CNN parameters are initialized and trained using PSO. Then, when PSO progress induction decelerates, the SGD algorithm is applied for few iterations. After few iterations, the process is switched to PSO and so on. In addition, PSO is consolidated by Genetic Algorithm (GA), which is exploited to stimulate particles and overcomes SGD lethargy. Moreover, unlike standard PSO, which requires a long time to reach the potent area, hybrid PSO provides fast and enhanced optimization [33, 34]. In this algorithm, it is endeavored to preserve the training CNN vital for the whole training period. Algorithm 1 shows the CNN training using the proposed hybrid training method. Algorithm 2 describes PSO alone as well.

Algorithm 1. CNN training using the proposed hybrid method

```
▪ Initialize parameters with different means
▪ $w^n = [w_1^n, w_2^n, \dots, w_p^n]$ and $b^n = [b_1^n, b_2^n, \dots, b_p^n]$
▪ Begin
▪     Repeat t
▪        For k ← 1 to P do
▪            Model evaluation $E_k = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{Ref} - \mathrm{CNN(output)})^2$
▪        End
▪        $\mathrm{pbest} = \min(E(p^{p(t)}), E(p^{p(t-1)}), E(p^{p(t-2)}) \dots E(p^{p(t-n)}))$
▪        $\mathrm{gbest} = \min(\mathrm{pbest}^1, \mathrm{pbest}^2, \mathrm{pbest}^3, \dots., \mathrm{pbest}^P)$
▪        Update (1) and (2)
▪       If (E) saturates for 3-8 iterations,
▪               go back to repeat t
▪         else
▪               Apply GA by choosing random particles from $P_1$ to $P_p$
▪         End
▪        Choosing 10% of the best particles
▪        For k ← 1 to n do
▪            Apply SGD with those particles $SGD(p_i)$
▪       End
▪    End
▪ Until the condition reached
```

Algorithm 2. PSO Algorithm

```
▪ randomly initialize CNN parameters
▪ Model Evaluation
▪ $gbest = \arg\min_j E(\Theta)\ ; j = 1..p$
▪ $\text{if}\ E(pbest^{p(t-1)}) > E(pbest^{p(t)}) => E(pbest^{p(t-1)}) = E(pbest^{p(t)})$
▪ $V_l^{p(t)} = V_l^{p(t-1)} + cr(pbest^p - \chi_l^{p(t-1)}) + cr(gbest - \chi_l^{p(t-1)})$
▪ $\chi_l^{p(t)} = \chi_l^{p(t-1)} + V_l^{p(t-1)}$
▪ until condition satisfied
```

V. Proposed Model Architecture

CNN generally consists of alternatives two main layers called convolution and max-pooling layer and end up with fully connected layer. All these layers are connected to each other with weights. However, there are many different other CNN architectures. In this study, the same structure proposed by Yann LeCun et al. [42] is used.

There are is of ambiguous steps, which need to be clarified such as how can the CNN parameters be encapsulated into particles? How do they cooperate with each other? How can it justify the best particles with the ensembles of swarm? To answer these questions, how the parameters of CNN are distributed. It is obvious that the weights and biases are constituent parameters of CNN. Therefore, in this work, the weights and bias are dismantled and encapsulated into vectors as shown below:

$$\mathbb{W} = \{w^1, w^2 \ldots..w^p \tag{3}$$

$$Ƥ = \{b^1, b^2 \ldots..b^p\} \tag{4}$$

$$w^n = \{w_1^n, w_2^n, \ldots, w_L^n\} \tag{5}$$

$$b^n = \{b_1^n, b_2^n, \ldots, b_L^n\} \qquad l = 1 \ldots L,\ \ n = 1 \ldots.P \tag{6}$$

where l is the layer index, L is the total number of layers, n is the particle n, P is the total number of particles, w_l^n is the weight parameters of layer l, and b_l^n is the bias parameters of the layer l. Finally the final total parameters of bias and weights are given by

$$W^n = \{\mathbb{W}, Ƥ\} \tag{7}$$

Fig. 2 shows the first convolution and max-pooling layers of CNN, and there are set of filters and each has dimensions $n \times n$ and can be vectored to be $1 \times n^2$. Thus, having m filters for lth layer, then the total weight parameters are $w_l^n = mn^2$. In addition, the total bias parameters for the given lth layer are $b_l^n = 1 \times m$.

Since there are P particles that will be trained, each one of them could be the best one among the swarm and can give an optimal solution. In order to justify which the best particle among swarm, the following notion is used: $w^* = \arg\min E(\Theta)$

where w^* is the best particle among swarm and E described below is the measured error between the reference and the model output.

$$E = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}(Ref_{ji} - d_{ji})^2 \tag{8}$$

Where N is the number of training samples, m is the number the output layers, Ref_{ji} is the reference, d_{ji} is the output of CNN. For clarification and showing the difference between the updating parameters using BP and PSO, fig. 2 shows the principle of how BP and PSO work. The figure has circles having functions g_n where $n = 1 \ldots L$ and L is the number of layers. The last layer has a function f.

It is noticeable that BP requires both forward and backward phases. In the forward phase, each activation function gives its response with respect to the input. In backward phase, the derivative is required with to respect to network parameters. PSO does not require any backward phase which can save vast expanse of work and time consumption because the forward phase is less problematical than backward phase.

The reason of why the second phase of network is not compulsory because the PSO algorithm depends on positions and velocities of the particles described in fig. 3. For instance, if there are P particles and the particle m is the best particle which satisfies (8), then the particle n can be updated according to (1) and (2).

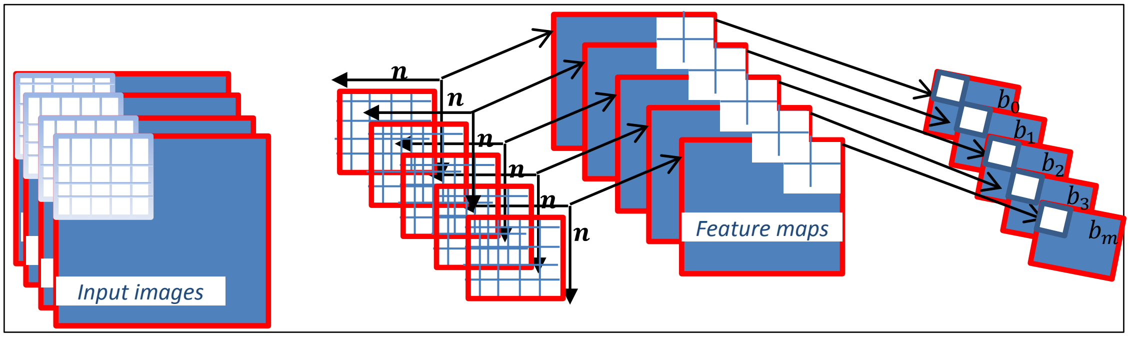


Fig. 1. Convolution and max-pooling layers of CNN network

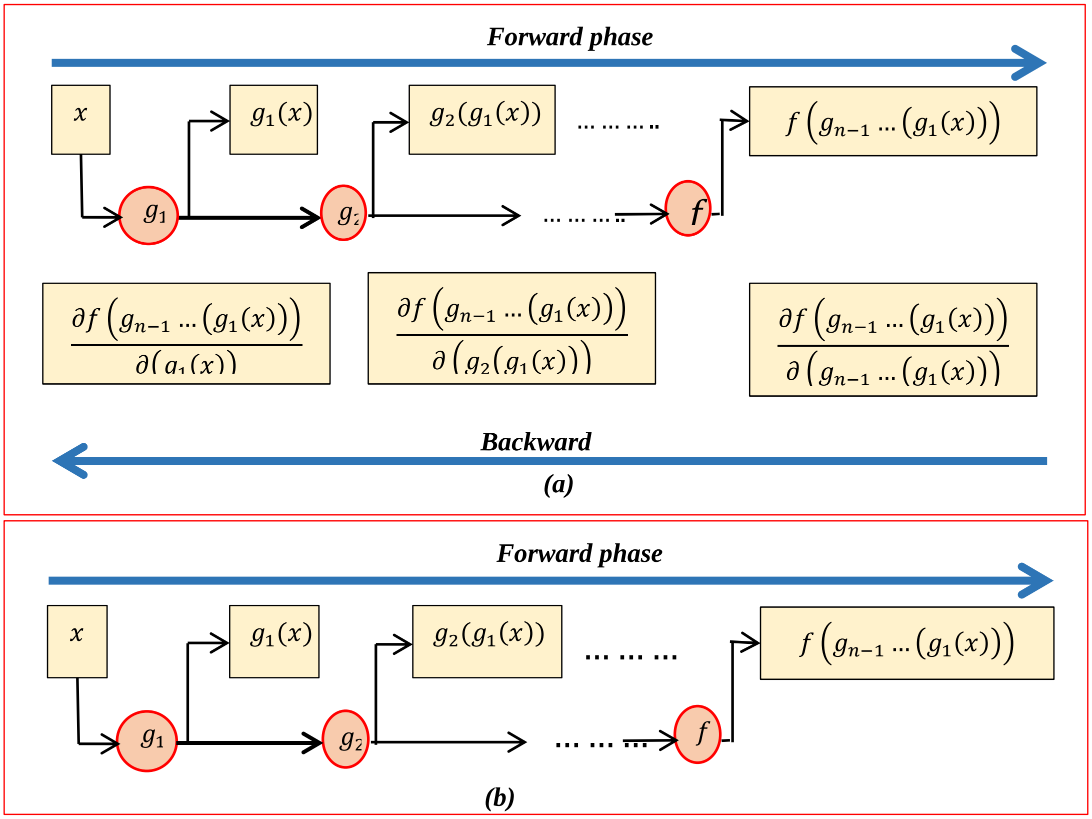

Forward phase
x
$g_1(x)$
$g_2(g_1(x))$
… … … …
$f\left(g_{n-1} \ldots (g_1(x))\right)$
g_1
g_2
… … … …
f
$\frac{\partial f\left(g_{n-1} \ldots (g_1(x))\right)}{\partial (g_1(x))}$
$\frac{\partial f\left(g_{n-1} \ldots (g_1(x))\right)}{\partial \left(g_2(g_1(x))\right)}$
$\frac{\partial f\left(g_{n-1} \ldots (g_1(x))\right)}{\partial \left(g_{n-1} \ldots (g_1(x))\right)}$
Backward
(a)
Forward phase
x
$g_1(x)$
$g_2(g_1(x))$
… … …
$f\left(g_{n-1} \ldots (g_1(x))\right)$
g_1
g_2
… … …
f
(b)

Fig. 2. The principle of working both algorithms (a) BP (b) PSO

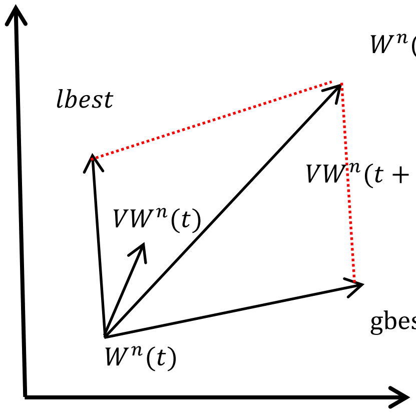


Fig. 3. The principle of parameters updating using PSO

Where $W^n(t)$ is the parameters mentioned in (7) at time t, gbest and lbest are mentioned in (1), $VW^n(t)$ is the velocity of particle n, and $W^n(t+1)$ is the next position for the CNN parameters of part n.

VI. Benchmark Experiments

A. Overview

The algorithm is evaluated on three benchmark datasets: MNIST [18], CIFAR-10 [43], and SVHN [44]. Samples for the datasets are shown in Fig 4. The CNN used in this work consists of alternative convolutional and max pooling layers. Fully connected layer is implemented on the top of the network. The architecture of CNN used for each dataset is dissimilar from each other. The number of particles is 25 and they are randomly initialized with different means and variances.

Fig. 4. Samples of (a) MNIST (b) CIFAR-10 (c) SVHN datasets

B. MNIST dataset

The MNIST [42] is a hand written digits 0-9. The dataset consists of 60000 samples. 50000 samples are used for training and the rest used for testing. All samples have the same size, which is 28x28 pixels. The pixels are scaled to be in [0, 1] before the training. There is no preprocessing or data augmentation utilized in this work. The CNN structure is 8C-8S-24C-24S-89C-90F-10F, where C stands for Convolution layer, S is for subsampling layer, and F is for full conned layer. In this dataset, the size of mini-batches is 128 images. The prosed hybrid PSO and SGD is exploited for training. At the beginning, the particles are trained using PSO only and Mean Square Error (MSE) mentioned in (3) is used as fitness assessment for the particles. The lowest MSE particle is the highest fitness is. In these experiments, MSE keeps dropping in few iterations and it saturates after that. To circumvent such margins, SGD and GA are launched when there is no further error dropping seen. SGD-GA is usually applied if error saturates between 5-8 iterations. Test accuracy is 0.9957 % for MNIST dataset. To best of the knowledge, this is the best reported result without preprocessing, augmentation, or dropout. A summary of the best published results on MNIST dataset is shown in Table I.

when a large dataset is used such as MNIT, which has 60000 gray images for training and 10000 for testing or CIFAR-10, which has 50000 color images for training and 10000 for testing with a mini-batch 128 sued, it influences PSO performance because it cannot choose the best particle which depends on only 128 images so local minimum occurs.

TABLE I. RESULTS ON MNIST DATASET

Method	Ref. #	Test Accuracy
Unsupervised Learning	[45]	0.64
What is the Best Multi-Stage	[22]	0.53
2-Layer CNN + 2-Layer NN	[46]	0.53
Stochastic Pooling	[46]	0.47
NIN + Dropout	[46]	0.47
Conv. maxout + Dropout	[47]	0.45
Hybrid PSO-SGD	**ours**	**0.43**

To tackle this problem, a hybrid training algorithm of PSO and SGD is used. Instead of than using single algorithm, by collaborating two algorithms with each other, a better performance is reached. Table I shows most of the state-of-the-art results on MNIST. A comparison is performed with only results that do not have preprocessing or they have the same architecture of CNN. It is clear that this work surpasses other works that do not use distortions or any preprocessing.

C. CIFAR-10 Dataset

The CIFAR-10 dataset consists of 10 classes of natural 32x32 RGB images with 50,000 for training and 10,000 for testing [19]. The CNN used for this dataset is described as: 12C-12S-48C-48S-89C-90F-10F, which is denoted to convolutional layer with 12 feature maps, subsampling layer, and a convolutional layer with 48 feature maps, subsampling layer, and a convolutional layer with 89 feature maps, and a fully connected output layer with 90 neurons, and a fully connected output layer with 10 outputs.

The subsampling layers have filters over non-overlapping region of size 2x2. The same steps are followed as in MNIST for training CNN. However, in this dataset, occurring in local optimum is faster than previous datasets so the number of times applying SGD is higher. It is determined that PSO-GA needs to be united by SGD as complicated dataset used such as CIFAR-10 because the MNIST dataset is easier for classification than CIFAR-10. Nevertheless, the benefit of using hybrid POS-SGD is still obtainable. The test accuracy gotten on this dataset is 82.41%.

From table II, it is evident that the proposed method surpasses the other state-of-the-art works. It is worth mentioning that only comparison with methods that use the same structure of CNN is considered. Any other techniques that can be very valuable for increasing accuracy such as dropout or drop-connect are not used. In this work, the same general structure proposed by Yann LeCun et al. [42] is used and only the training algorithm is replaced but the same configuration of CNN is kept.

TABLE II. TEST SET ACCURACY RATES ON CIFAR-10 DATASET

Method	Reference #	Accuracy
Tiled CNN	[48]	73.10
Improved LCC	[49]	74.50
KDES-A	[50]	76.00
PCANet-2 (combined)	[51]	78.67
PCANet-2	[51]	77.14
K-means (Triangle, 4000 features)	[52]	79.60
Cuda-convnet2	[53]	82.00
Hybrid PSO-SGD	**[ours]**	**82.41**

D. SVHN Dataset

The last experiment is assessed on the street view house numbers (SVHN). The dataset consists of 604,388 samples (training and extra set) and 26,032 samples as test images. In addition, each the dataset is color images and the size of each sample is 23x32 pixels. Following [1, 2], 400 samples per class from the training set and 200 images per class from extra set are selected to implement validation set. The task in this dataset is to classify the digit in the center of each image. Preprocessing local contrast normalization is used following Goodfellow et al. [6]. In addition, the same CNN assembly and parameters setting are used as CIFAR-10. The test error obtained is 2.48%. The result is shown in Table III.

TABLE III. TEST ERROR RATES ON SVHN DATASET

Method	Reference #	Test Error %
Multi-Stage Conv. Net + 2-layer Classifier	[44]	5.03
Multi-Stage Conv. Net + 2-layer Classifier + padding	[44]	4.90
Maxout Networks	[16]	2.47
Hybrid PSO-SGD	**[ours]**	**2.48**

Again Maxout Networks is used in very large CNN implementations because it is implemented over Krizhevsky et al. [31] code. However, a conventional CNN is used instead. In addition, a leveraging PSO algorithm is used in this work which is faster than SGD

VII. CONCLUSION

In this work, a new hybrid training process is proposed and demonstrated called Particle Swam Optimization- Stochastic Gradient Decent (PSO-SGD) algorithm, for training Convolution Neural Network (CNN). It is established that the algorithm is well suited for achieving nontrivial results on different datasets and surprisingly achieving state-of-the-art on these datasets. The proposed algorithm is a proficient method for training because it combines both PSO and SGD in an innovative fashion. Analysis also shows that the proposed method is superior on three different benchmark datasets. The hybrid training method avoids occurring in local optimum and premature saturation inspired by using single algorithm. Additionally, it preserves the training vital for the whole training period and restrains the lethargy inherited by a monocular algorithm.

VIII. FUTURE WORK

In future more influential parameters will be explored. There are more parameters that can influent model accuracy will be investigated in the future work. Deeper analysis and more challenge datasets such as ImageNet also will be as a part of the future work. Also reporting time consumption and how fast execution time for training and testing will be consider endeavoring to reach real time execution.

REFERENCES

[1] S. L. Phung and A. Bouzerdoum, ”A pyramidal neural network for visual pattern recognition,” IEEE Transactions on Neural Networks, vol. 27, no. 1, pp. 329343, 2007

[2] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, Lior Wolf , “DeepFace: Closing the Gap to Human-Level

[3] Min Lin, Qiang Chen, and Shuicheng Yan “Network In Network” arXiv 1312.4400v3,4 Mar 2014

[4] Dan Cires¸an, Ueli Meier and Jurgen Schmidhuber, “Multi-column Deep Neural Networks for Image Classification” CVPR 2012

[5] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid, “Convolutional Kernel Networks” arXiv 14 Nov 2014 .

[6] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, Yoshua Bengio “Maxout Networks” ICML 2013

[7] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov, “Scalable Object Detection using Deep Neural Networks” CVPR, 2014.

[8] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan ,” Object Detection with Discriminatively Trained Part Based Models”

[9] R. Fergus, P. Perona, and A. Zisserman, “Object class recognition by unsupervised scale-invariant learning,” in IEEE Conference on Computer Vision and Pattern Recognition, 2003

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013. 1, 2.

[11] B. Leibe, A. Leonardis, and B. Schiele, “Robust object detection with interleaved categorization and segmentation,” International Journal of Computer Vision, vol. 77, no. 1, pp. 259–289, 2008.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006. 1, 2, 5, 6, 7

[13] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, CalTech, 2007.

[14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. arXiv preprint arXiv:1403.6382, 2014. 1, 2.

[15] Y. Lin, T. Liu, and C. Fuh. Local ensemble kernel learning for object category recognition. In CVPR, 2007

[16] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In ECCV, 2004.

[17] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In ICCV, 2005.

[18] Fabien Lauer, Ching Y. Suen, and G´erard Bloch “A trainable feature extractor for handwritten digit recognition” Journal Pattern Recognition 2007.

[19] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, Zhuowen Tu, “ Deeply-Supervised Nets “NIPS 2014

[20] M. Fischler and R. Elschlager, “The representation and matching of pictorial structures,” IEEE Transactions on Computer, vol. 22, no. 1, 1973.

[21] Kaiming, He and Xiangyu, Zhang and Shaoqing, Ren and Jian Sun “Spatial pyramid pooling in deep convolutional networks for visual recognition” European Conference on Computer Vision, 2014

[22] Kevin Jarrett, Koray Kavukcuoglu, Marc’Aurelio Ranzato and Yann LeCun ” What is the Best Multi-Stage Architecture for Object Recognition?” ICCV’09, IEEE, 2009.

[23] Ross Girshick, “Fast R-CNN “arXiv preprint arXiv:1504.08083, 2015

[24] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In ICCV, 2013. 8

[25] Karen Simonyan and Andrew Zisserman "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION" arXiv:1409.1556v5 [cs.CV] 23 Dec 2014.

[26] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. Internatinal Conference on Learning Representation, 2013. 2.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR, abs/1311.2524, 2013. 4

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 2.

[29] L. N. Clement Farabet, Camille Couprie and Y. LeCun. Learning hierarchical features for scene labeling. PAMI, 35(8), 2013. 1, 2

[30] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations".

[31] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (NIPS'2012). 2012.

[32] Frans van den Bergh and Andries P. Engelbrecht "A Cooperative Approach to Particle Swarm Optimization" IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION,JUNE 2004,VOL. 8, NO. 3.

[33] Shigenori Naka et al."A Hybrid Particle Swarm Optimization for Distribution State Estimation", IEEE TRANSACTIONS ON POWER SYSTEMS, FEBRUARY 2003, VOL. 18, NO. 1.

[34] X. Wang" A Hybrid Particle Swarm Optimization Method", IEEE International Conference on Systems, Man, and Cybernetics, October, pp. 4151-4157, 2006.

[35] Jiang Du et al. , "Improved PSO Algorithm and Its Application in Optimal Design for Rectifier Transformer", IEEE International conference on Intellignet and Integrated System, pp. 605-608, 2010,

[36] Yi-Xiong Jin et al. "Local Optimum Embranchment Based Convergence Guarantee Particle Swarm Optimization and Its Application in Transmission Network Planning", IEEE/PES Transmission and Distribution Conference & Exhibition: Asia and Pacific Dalian, China, pp. 1-6, 2005.

[37] Shigenori Naka et al."A Hybrid Particle Swarm Optimization for Distribution State Estimation", IEEE TRANSACTIONS ON POWER SYSTEMS, FEBRUARY 2003, VOL. 18, NO. 1.

[38] Zongmei Zhang, Zhifang Sun, and Hiroki TAMURA, "Local Linear Wavelet Neural Network with Weight Perturbation Technique for Time Series Prediction", IEEE international Conference on Computer Science and Software Engineering, pp. 789-801,2008.

[39] Zhaohe Huang et al, "A Particle Swarm Optimization Algorithm for Hybrid Wireless Sensor Networks Coverage", IEEE Symposium on Electrical & Electronics Engineering (EEESYM), pp. 630-632, 2012.

[40] Esmaeil Hadavandi, Arash Ghanbari, and Salman Abbasian-Naghneh, "Developing a Time Series Model Based On Particle Swarm Optimization for Gold Price Forecasting", IEEE Third International Conference on Business Intelligence and Financial Engineering, pp. 337-340, 2010.

[41] Mohammad Yunus Ali and Kaamran Raahemifar, "Reactive Power Optimization Based on Hybrid Particle Swarm Optimization Algorithm", IEEE Canadian Conference on Electrical and Computer Engineering, pp. 1-5, 2012.

[42] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffiner "Gradient-Based Leaning Applied to Document Recognition" Proc of the IEEE, November 1998.

[43] A. Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Computer Science Department, University of Toronto, 2009. 1, 6.

[44] P. Sermanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit classification. In ICPR, 2012

[45] Marc'Aurelio Ranzato, Fu-Jie Huang, Y-Lan Boureau, Yann LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition" CVPR, 2007

[46] Matthew D Zeiler and Rob Fergus. Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557, 2013.

[47] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. arXiv preprint arXiv:1302.4389, 2013.

[48] Q. V. Le, J. Ngiam, Z. Chen, D. Chia, P. W. Koh, and A. Y. Ng, "Tiled convolutional neural networks," in NIPS, 2010.

[49] K. Yu and T. Zhang, "Improved local coordinate coding using local tangents," in ICML, 2010.

[50] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in NIPS, 2010

[51] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma "PCANet: A Simple Deep Learning Baseline for Image Classification? "arXiv:1404.3606v2 [cs.CV] 28 Aug 2014.

[52] Adam Coates, Honglak Lee, and Andrew Y. Ng "An Analysis of Single-Layer Networks in Unsupervised Feature Learning" In AISTATS 14, 2011.

[53] A. Krizhevsky, "cuda-convnet," http://code.google.com/p/ cuda-convnet/, July 18, 2014.

Cloud Management and Governance: Adapting IT Outsourcing to External Provision of Cloud-Based IT Services

Dr. Victoriano Valencia García
Computer Management Technician
and Researcher at Alcalá University
Madrid, Spain

Dr. Eugenio J. Fernández Vicente
Professor at Computer Science Dept.
Alcalá University
Madrid, Spain

Dr. Luis Usero Aragonés
Professor at Computer Science Dept.
Alcalá University
Madrid, Spain

***Abstract*—Outsourcing is a strategic option which complements IT services provided internally in organizations. The maturity model for IT service outsourcing (henceforth MM-2GES) is a new holistic maturity model based on standards ISO/IEC 20000 and ISO/IEC 38500, and the frameworks and best practices of ITIL and COBIT, with a specific focus on IT outsourcing. MM-2GES allows independent validation, practical application, and an effective transition to a model of good governance and management of outsourced IT services.**

Cloud computing is a new model for provisioning and consuming IT services on a need and pay-per-use basis. This model allows the IT systems to be more agile and flexible. The external provision of cloud-based services as part of Cloud computing appears as an evolution of traditional outsourcing, due to the emerging technologies related to the provision of IT services. As a result of technological developments, traditional outsourcing and external provision of cloud-based services, share common characteristics, but there are also some differences.

This paper adapts MM-2GES to external provision of cloud-based services, from the point of view of the customer. This way, the applicability of the model can be implemented in organizations that have both models traditional IT outsourcing and cloud-based services provided externally, in order to achieve excellence in governance and management of all kind of IT services provided externally to organizations.

Keywords—Cloud computing; IT governance; IT management; Outsourcing; IT service; Maturity model

I. INTRODUCTION

One thing to change about Information Technologies (henceforth IT) at university level is the deeply rooted approach which exists, or which used to exist, called infrastructure management. This kind of management has evolved into a governance and management model more in line with the times, which is a professional management of services offered to the university community [4]. It is for this reason that in recent years a set of methodologies, best practices and standards, such as ITIL, ISO 20000, ISO 38500 and COBIT, have been developed to facilitate IT governance and management in a more effective and efficient way.

These methodologies, which are appropriate and necessary to move from infrastructure management to service management, see a lack of academic research. For that reason it is inadvisable to use these frameworks on their own, and it is advisable to consider other existing frameworks in order to extract the best from each for university level [4].

IT services have implications for business and innovation processes and may be a determinant in their evolution. The organization of these services, their status within the organization of the university, and their relationships with other management departments and new technologies is therefore vital. At present, the degree of involvement, the volume of services offered, and the participation or external alliances with partner companies through outsourcing, that Gottschalk and Solli-Saether [5] defined as the "practice of turning over all or part of an organization's IT function to an IT vendor", are of special interest.

Currently, and in the years to come, organizations that achieve success are and will be those who recognize the benefits of information technology and make use of it to boost their core businesses in an effective strategic alignment, where delivery of value, technology, risk management, resource management, and performance measurement of resources are the pillars of success.

It is necessary to apply the above-mentioned practices through a framework and process to present the activities in a manageable and logical structure. Good practice should be more strongly focused on control and less on execution. They should help optimize IT investments and ensure optimal service delivery.

It is clear that ITs have become ubiquitous in almost all organizations, institutions and companies, regardless of the sector to which they belong. Hence, effective and efficient IT management to facilitate optimal results is necessarily essential.

Furthermore, in this environment of total IT dependency in organizations using ITs for the management, development and communication of intangible assets, such as information and knowledge [7], organizations become successful if these assets are reliable, accurate, safe and delivered to the right person at the right time and place, according to ITGI. Also, knowledge integration mechanisms is important in helping knowledge utilization in client firms [10].

In short, Fernández [4] proposes that the proper administration of IT will add value to the organization, regardless of its sector (whether social, economic or academic) and will assist it in achieving its objectives and minimizing risk.

Given the importance of proper management of IT, the search for solutions to the alignment of IT with the core business of organizations has accelerated in recent years. The use of suitable metrics or indicators for measurement and valuation, generate confidence in the management teams. This will ensure that investment in ITs generates the corresponding business value with minimal risk [4].

The above solutions are models of good practice, metrics, standards and methodologies that enable organizations to properly manage ITs. And public universities are not outside these organizations, though they are not ahead. In addition, interest in adopting models of governance and management of appropriate ITs is not as high as it should be.

Two of the factors through which IT best practices have become important is, the selection of appropriate service providers and the management of outsourcing and procurement of IT services.

IT outsourcing has brought potential benefits in addition to many examples of the great organizational losses associated with this practice. Even with awareness of the potential for failure, the IT outsourcing industry continues to grow, as organizations communicate their desire to engage in IT outsourcing and their determination to decipher a method that enables successful IT outsourcing relationships [9].

On the other hand, Cloud computing is the latest trend to outsource some or complete IT operations to run a business from the public cloud that provides a flexible highly scalable technology platform for an organization’s business operations [1]. In addition, Cloud computing poses serious challenges to traditional business process outsourcing and have a profound impact on how IT outsourcing is done [11].

Cloud computing represents a fundamental shift in how organizations pay for and access IT services. It has created new opportunities for IT service providers and the outsourcing vendors. Cloud computing will have significant impact on outsourcing vendors, who must adopt new strategies to include Cloud services as part of their offerings to keep up with profound changes in the IT service industry. They should experiment with Cloud services and understand which models are suitable for their clients. This will help them to identify new business opportunities that arise from cloud computing. In addition, the deployment of new innovative Cloud services with attractive business models will lead to high level of customer satisfaction and unprecedented adoption of Cloud services in the organizations [3].

In the following study, a new holistic maturity model with a specific focus on IT outsourcing, is adapted to external provision of cloud-based services, taking into account the general characteristics that define IT outsourcing and external provision of cloud-based services, but also bearing in mind the nuances outlined on service models and deployment models. After analysing all the differences between traditional IT service outsourcing and IT services provided externally from the cloud, the justified adjustments to be made in the indicators of the new holistic maturity model are shown in order to adapt it to the external provision of cloud-based IT services, from the point of view of the customer.

II. MM-2GES AND CLOUD-BASED IT SERVICES PROVIDED EXTERNALLY

MM-2GES is a new holistic maturity model based on standards ISO/IEC 20000 and ISO/IEC 38500, and the frameworks and best practices of ITIL and COBIT, with a specific focus on IT outsourcing. The model allows independent validation, practical application, and an effective transition to a model of good governance and management of outsourced IT services.

In order to design the model, we studied in detail every reference on the provision of IT services that there is in the ISO 20000 and ISO 38500 standards and ITIL v3 and COBIT methodologies, with a specific focus on IT outsourcing.

The model follows a stage structure and has two major components: maturity level and concept. Each maturity level is determined by a number of concepts common to all levels. Each concept is defined by a number of features that specify the key practices which, when performed, can help organizations meet the objectives of a particular maturity level. These characteristics become indicators, which, when measured, determine the maturity level.

MM-2GES defines five maturity levels: initial or improvised; repeatable or intuitive; defined; managed and measurable; and optimized. The measurement tools of the model is the subject of a paper published in an International Conference.1

The model proposes that organizations under study should ascend from one level of maturity to the next without skipping any intermediate level. In practice, organizations can accomplish specific practices in upper levels. However, this does not mean they can skip levels, since optimum results are unlikely if practices in lower levels go unfulfilled.

The applicability of the model and the measurement tools, allow organizations to meet the goal of effective transition to a model of good governance and good management of outsourced IT services.

On the other hand, the external provision of cloud-based services appears as a natural evolution of traditional outsourcing, due to the emergence of emerging technologies related to the provision of IT services. As a result of technological developments, traditional outsourcing and external provision of cloud-based services, share common characteristics such as reduced costs, increased flexibility, simplified IT and release of resources, among others. But the external provision of cloud-based services has some characteristics that are distinct from traditional outsourcing, from the point of view of the customer, such as shorter

[1] Cf. Valencia, V., Usero, L., & Fernández, E., “Measurement Tools of the Maturity Model for IT Service Outsourcing in Higher Education Institutions”, Proceedings of the 13th International Conference on Information Systems Design and Technology, Jan 13-14, 2015, Zurich, Switzerland, 9 (01) Part V, pp. 667-676

contracts, more transparent costs, less project management, less government and less coordination, among others.

MM-2GES is a good basis in order to use in this new scenario and new business model based on external provision of cloud computing services. The model has been designed to be applied in any scenario where an organization has the ability to hire IT services to external service providers. However, there are some significant differences between traditional IT outsourcing and external provision of services based in the cloud. Therefore, the model allows some adjustments to be made in the metrics tables where each characteristic in the model is rated. The adjustments in the model must be made depending on the differences mentioned above.

In addition, when discussing cloud computing we must take into account there are three different service models, each of which has specific characteristics, and also there are four deployment models with specific characteristics. Thus, when discussing the external provision of services in the cloud, the characteristics of services depend on the type, design and nature of the service, studied in a personalized way.

III. Differences Between Traditional It Service Outsourcing and Cloud-Based It Services Provided Externally

In the following analysis, MM-2GES is adapted to external provision of cloud-based services. These kinds of services are analysed from a general and theoretical angle, taking into account the general characteristics that define them, but also bearing in mind the nuances outlined on service models and deployment models.

Therefore, the differences between traditional IT service outsourcing and IT services provided externally from the cloud are the following, extracted from different publications [2], [3], [6], [8]:

- [D1] No advance costs. Simpler and more transparent costs;
- [D2] Shorter, less complex and more important contracts;
- [D3] Rapid scaling on demand. Greater flexibility with respect to the increase and decrease of IT resources needed due to the existence of services and infrastructure deployed in the cloud;
- [D4] Less customization of services and greater difficulty of integrating legacy systems;
- [D5] Less project management, less government and less coordination. Less interaction;
- [D6] Few legal guarantees on security and data privacy;
- [D7] Greater uncertainty about business continuity;
- [D8] Greater uncertainty and importance of service availability; and
- [D9] Less guarantee of getting a certain performance.

Taking into account the differences between IT outsourcing and the external provision of cloud-based services, it is necessary to make a number of adjustments in MM-2GES. These adjustments will allow to apply the maturity model in an environment where external cloud-based services are provided. This way, a new characteristic that define both traditional outsourcing and cloud computing will be added to the maturity model, flexibility. This requires identifying which concepts of MM-2GES would be affected by the differences between traditional IT outsourcing and the external provision of cloud-based services. Table I shows the concepts affected by the differences. The first column of the table shows all key areas that form the basis of the maturity model designed for IT outsourcing. The second and successive columns show the differences between IT outsourcing and external provision of cloud-based services.

TABLE I. Key Areas Affected by the Differences between IT Outsourcing and External Provision of Cloud-Based Services

Key areas or determinants	**Differences between IT outsourcing and external provision of cloud-based services**								
	D1	***D2***	***D3***	***D4***	***D5***	***D6***	***D7***	***D8***	***D9***
Formal Agreement		X	X						
Service Measurement									X
Quality Management									
Monitoring and Adjustments					X				X
Alignment IT-Business				X					
IT Governance Structure					X				
Service Level Agreement (SLA)		X	X					X	X
IT Service Registration									
Incident and Problem Management									
Changes									
Testing and Deployment									
Control of External Providers								X	
Business Risk							X		
Financial Management	X								
Legislation						X			
Demand and Capacity Management			X				X		
Formal Agreement Management		X							
Knowledge Management									
Guidelines on outsourcing an IT service (life cycle)	X	X							

The concepts that are not altered by the differences between IT outsourcing and external provision of cloud-based services are the following:

- Quality Management;
- IT Service Registration;
- Incident and Problem Management;
- Changes;
- Testing and Deployment; and
- Knowledge Management.

The concepts altered by the differences between IT outsourcing and external provision of cloud-based services are the following:

- Formal Agreement;
- Service Measurement;
- Monitoring and Adjustments;
- Alignment IT-Business;
- IT Governance Structure;
- Service Level Agreement (SLA);
- Control of External Providers;
- Business Risk;
- Financial Management;
- Legislation;
- Demand and Capacity Management;
- Formal Agreement Management;
- Knowledge Management; and
- Guidelines on outsourcing an IT service (life cycle).

IV. Adjustments in Order to Adapt Mm-2ges to the Cloud

After analysing all the differences, extracted from different publications [2], [3], [6], [8], between traditional IT service outsourcing and IT services provided externally from the cloud, the adjustments to be made in the indicators of the model in order to adapt it to the external provision of cloud-based IT services, are the following:

A. *Formal Agreement*

Differences between external provision of cloud-based services and IT outsourcing: [D2, D3]

Because contracts in external provision of cloud-based services are in general shorter in duration, less complex and more important, in addition to the scaling of IT resources potentially demanded by organizations is more demanding, formal agreements signed by the customer and cloud supplier should be more flexible. Therefore, the review of the formal agreements in external provision of cloud-based services should be more frequent at predefined intervals. This circumstance would allow the ability to align quick scaling of IT resources and the provisions of the signed formal agreements.

Affected indicators: ACF2e, ACF4.

B. *Service Measurement*

Differences between external provision of cloud-based services and IT outsourcing: [D9]

Because the external provision of cloud-based IT services share the same infrastructure for all clients who hire these services, it is more difficult to get a certain performance due to the use the rest of the customers can make of shared services. This way, customers can achieve high availability of services and have poor performance at the same time.

As a result, measuring the performance of cloud-based services provided externally is essential to meet the expectations and business needs. In addition, measurements allow early detection of potential problems in services, according to Cobit. Therefore, it would be necessary to have procedures for the measurement of cloud-based servicies in order to assess their performance. These procedures would facilitate the achievement of the objectives in organizations.

Affected indicators: MED1.

C. *Monitoring and Adjustments*

Differences between external provision of cloud-based services and IT outsourcing: [D5, D9]

Although it is needed less management and less government in external provision of cloud-based services, and the coordination between organizations and cloud providers is less necessary, all model indicators that define the concept of monitoring and adjustments are not altered, but the existence of the KPI indicators needed to assess the performance of services so that they can meet the needs and expectations of the business.

Therefore, the characteristics that define the indicators for monitoring and adjustments, and that they are important to be kept in the external provision of cloud-based services in the same way as in IT outsourcing, are the following:

- Monitoring IT services;
- KPIs and KGIs affect the penalties, contracting and negotiation of IT services;
- Meeting legal and operational requirements; and
- Corrective actions as a result of monitoring.

Affected indicators: MON2.

D. *Governance Structure*

Differences between external provision of cloud-based services and IT outsourcing: [D5]

In an environment of cloud-based services provided externally is needed less government by client organizations, but the IT governance structure established in the model remains the same because their functions are still valid, and therefore the following committees and commissions:

- IT Strategy Committee;
- Audit Committee;
- IT Steering Committee;
- Commission on Technology and / or IT Architecture;
- Services Commission; and
- Projects Office.

E. Service Level Agreement (SLA)

Differences between external provision of cloud-based services and IT outsourcing: [D2, D3, D8, D9]

Due to the fact that cloud-based services provided externally are under the tutelage of agreements or contracts shorter in time, and demand of scaling IT resources are more demanding, SLAs should be reviewed more often as happens with formal agreements. In addition, is important to keep availability in SLAs in the same degree as in traditional outsourcing, require greater compliance with the performance of cloud-based services provided externally due to the possible scaling up or down of services, and reduce security requirements due to the inability to obtain a full security commitment by cloud providers, due to the intrinsic cloud configuration.

Affected indicators: SLA2a, SLA2b, SLA2g, SLA3.

F. Control of External Providers

Differences between external provision of cloud-based services and IT outsourcing: [D8]

In an environment of external provision of cloud-based services, the availability of services becomes more important because it creates more uncertainty having the infrastructure that supports services in locations sometimes physically unknown. Cobit indicates the need to assign responsibility for managing external suppliers and the quality of services provided. In addition, it specifies that the external provider is subject to periodic independent reviews to feedback its performance in order to improve its service delivery.

Therefore, is necessary to establish a control on external suppliers by conducting independent audits and security reports. The latter allow, among others things, to monitor and ensure confidentiality, integrity and availability of information about the services provided by external suppliers, minimizing risk.

In addition, is important to keep the characteristics that define the indicators on the control of external suppliers in external provision of cloud-based services to the same degree as in outsourcing.

Affected indicators: CPE2.

G. Business Risk

Differences between external provision of cloud-based services and IT outsourcing: [D7]

The external provision of cloud-based services needs to improve important aspects such as security of information managed. This formula creates uncertainty and distrust in organizations because of potential risk in the business. This circumstance forces to adjust the indicators that define the concept of business risk in the model, and thus mitigate the uncertainty created.

Therefore, it is important to assess the risk, have a contingency plan that is regularly reviewed as well as the skills and capabilities of the cloud provider, as part of the business risk management. Strict compliance with service levels offered, quality of services and business continuity, reduce business risk.

Affected indicators: RIN1, RIN2, RIN3, RIN4.

H. Financial Management

Differences between external provision of cloud-based services and IT outsourcing: [D1]

Costs of services provided externally in the cloud are more transparent than those arising when external supplier provides services using traditional outsourcing. In addition, there are no costs in advance in external provision of cloud-based services. Despite making clear the costs, it would be desirable to maintain a commission in financial management. This commission would provide vital information for IT management in order to ensure external provision of cloud-based services efficiently and profitable.

I. Legislation

Differences between external provision of cloud-based services and IT outsourcing: [D6]

It is well known that one of the reasons that keep many organizations from taking the step to the cloud, is the lack of legal guarantees on security and privacy of data providers manage in the cloud. Providing confidential information to another organization is always delicate and produces distrust. However, organizations that offer cloud services are aware they put their reputation at risk; that is why they try to dispel mistrust founded. In addition, data provided by cloud services can be accessed from any location. Therefore, there is always the possibility of compromising customer privacy.

Taking all the above into account, it is necessary to relax the enforcement of rules, laws, decrees, directives and decisions about data protection, data processing, processing site, clauses for the transfer of data and contractual clauses for the transfer of personal data to third countries. This involves to cover all levels of the maturity model.

Affected indicators: LEG1.

J. Demand and Capacity Management

Differences between external provision of cloud-based services and IT outsourcing: [D3, D7]

Managing demand of cloud services is less complex than in traditional outsourcing, because the scaling is faster and more flexible, and the availability of IT resources required are provided by cloud providers almost immediately. This does not prevent uncertainty in business continuity from increasing. This is due to the inability to act directly on the IT infrastructure that supports services in the event of a disruption produced by a client request to increase or decrease the capacity in cloud services.

Affected indicators: GDC1, GDC2, GDC3.

K. Formal Agreement Management

Differences between external provision of cloud-based services and IT outsourcing: [D2]

The formal agreements signed with cloud-based services suppliers are generally shorter in duration than the formal agreements signed with traditional outsourcers. This circumstance requires the availability of a management system of formal agreements signed with cloud providers to facilitate getting a consistent quality service at a competitive price, plus an extension, renewal and/or renegotiation of formal agreements.

In addition, the management system of formal agreements signed with cloud providers should be integrated into a configuration management system, which allows to have a clear knowledge and control of the infrastructure, the relationships between configuration items (ECs) that make up the infrastructure and support services, and life cycle of the ECs.

Finally, it would necessary to create a new role in the organization named contract manager or similar, responsible for managing formal agreements or contracts with external suppliers through the management system of formal agreements. This circumstance requires the creation of a new indicator in the metrics table (see Table II).

Affected indicators: GAF1.

New indicator: GAF4.

L. Guidelines on outsourcing an IT service (life cycle)

Differences between external provision of cloud-based services and IT outsourcing: [D1, D2]

When an organization faces the compromising decision to outsource to the cloud an IT service that was previously provided internally, the organization should analyze comprehensively all the circumstances related to that decision from an economic, technical and regulatory compliance point of view before making a final decision. Services provided externally in the cloud have the added difficulty for organizations of not knowing the physical location of the infrastructure that supports the services. This implies a greater difficulty in applying enforcement regulations to information that cloud providers manage.

Costs of services provided externally in the cloud are more transparent than those arising when external supplier provides services using traditional outsourcing. In addition, there are no costs in advance in external provision of cloud-based services. This circumstance facilitates to carry out an economic study to compare costs of cloud-based services with costs of the same services provided internally, including technical staff, equipment (processing and storage) and infrastructure costs.

Furthermore, it is necessary to carry out a technical study using proper indicators, with a specific focus on availability, continuity and capability of the service provided internally by the organization, if any, and that is meant to outsource to the cloud.

It is also necessary to assess the service demand required by the organization. In the assessment is included the following: percentage of required availability, quality, degree of continuity and capacity, not to mention in this assessment current and future resources available internally.

Once you have completed the above studies and the assessment is favorable to outsource the service studied to the cloud, it's time to explore the market of service providers in the cloud in order to contrast the service provided internally and the cloud-based service provided externally. Bids received should be renegotiated downward economically, while maintaining all the benefits of the service.

Affected indicators: PAS1a, PAS1b.

Finally, Table II shows the necessary adjustments to be made in the indicators of the model in order to adapt it to the external provision of cloud-based services. Indicators to be adjusted are highlighted in yellow. Indicator levels subject to adjustment are highlighted in green. New indicators, in this case there is only one (Contract Manager), are highlighted in green.

TABLE II. ADJUSTMENTS IN METRICS TABLE AND QUESTIONNAIRE

Level	Code - Indicator - Question of Questionnaire	Source
Concept: Formal Agreement: contract, agreement or similar (FA)		ISO 20000, Cobit 4.1, ITIL
4 > 2	**FA4** - Revision frequency of FAs - Formal agreements are reviewed periodically at predefined intervals	ISO 20000
Concept: Service Measurement (MED)		Cobit
3 > 2	**MED1** - Measurement procedures - There are clear procedures for the measurement (quality, performance, risks) of outsourced IT services	Cobit
Concept: Monitoring and Adjustments (MON)		Cobit & ITIL
3 > 2	**MON2** - There are optimized key performance indicators (KPIs) and key goal indicators (KGIs)	Cobit & ITIL
Concept: Service Level Agreement (SLA)		ISO 20000 e ITIL
2	**SLA1** - SLA – There is an SLA for each outsourced IT service provided by the service provider	ISO 20000 e ITIL
	SLA2 - Elements of SLA - SLAs include:	
2	**SLA2a** - Service availability	
5 > 4	**SLA2b** - Service performance	Myself
3	**SLA2c** - Penalties for breach of SLA	
2	**SLA2d** - Responsabilities of the parties	
3	**SLA2e** - Recovery Times	
4	**SLA2f** - Quality Levels	
4 > 5	**SLA2g** - Security requirements	
3 > 2	**SLA3** - Frequency reviewing of SLA - SLAs are reviewed periodically at predefined intervals	ISO 20000 and myself
Concept: Business risk (RIN)		ITIL & myself
4 > 3	**RIN1** - Contingency Plan (CP) - There is a backtracking contingency plan on outsourced	ITIL & myself

	IT services that support the core business, as part of the business continuity plan	
4 > 3	**RIN2** - Review CP - The contingency plan is reviewed periodically	ITIL & myself
3 > 2	**RIN3** - Business Risk Assessment - The business risk associated with outsourced IT services is assessed by the CIO or similar figure and reported to the appropriate IT governance body	Myself
4 > 3	**RIN4** - Skills and capabilities of the supplier - The skills and capabilities of the external provider are checked continuously, as part of the business risk management. I.e., strict compliance with the signed agreements is checked: the service levels offered, the quality of services and business continuity	ITIL
Concept: Legislation (LEG)		LOPD
123 > 12345	**LEG1** - Compliance with legislation - Degree of compliance with rules, laws, decrees, European and national directives and decisions about data protection, data processing location, clauses for the transfer of data and standard contractual clauses for the transfer of personal data to third countries, etc…	LOPD & myself
Concept: Demand and capacity management (DCM)		ITIL
4 > 3	**GDC1** - Demand management process - Demand management process (ITIL strategy phase) is implemented in order to regulate the demand for outsourced IT services	ITIL
4 > 3	**GDC2** - Capacity management process - Capacity management process (ITIL design phase) is implemented in order to ensure that IT capacity meets the current and future needs of outsourced IT services	ITIL
45 > 34	**GDC3** - Demand-capacity gearing - To what extent are demand management and capacity management (always justifiable in terms of costs) of outsourced IT services geared to each other?	ITIL & myself
Concept: Formal Agreement Management (contracts, agreements or similar) (GAF)		ISO 20000 & ITIL
4 > 3	**GAF1** - Formal agreement management system (AMS) – There is a management system of formal agreements signed with external IT providers in order to achieve a consistent quality service at a competitive price	ISO 20000 & ITIL
3	**GAF4** - Contract Manager - There is a contract manager, responsible for managing the agreements signed with external providers	Myself
Concept: Guidelines on outsourcing an IT service (PAS)		TFM (Myself)
3 > 4	**PAS1a** - Legislation study - A comprehensive legislation study on rules, laws, decrees, European and national directives and decisions, which the external provider must comply with when providing the service, has been carried out	TFM
3 > 2	**PAS1b** - Economic study of costs - A comprehensive economic study on the annual costs of the services supported internally, including technical staff, equipment (processing and storage) and infrastructure	TFM

The necessary adjustments to be made into the metrics table (see Table II) allow to adapt the maturity model for IT service outsourcing to external provision of cloud-based services. Then is when the applicability of the model can be implemented in organizations that have cloud-based services provided externally. In order to apply the new model we must use the same tools used in the maturity model for IT service outsourcing. The tools are the following: a questionnaire, metrics table, and continuous improvement plan tables as part of the measurement process. Therefore, the procedure to follow is to put into practice the measurement process by implementing the continuous improvement plan, and using the assessment tool, which is composed of the questionnaire results, metrics table and continuous improvement plan tables.

V. Conclusions

Cloud computing is a new model for provisioning and consuming IT services on a need and pay-per-use basis. This new paradigm of information technology is a model that enables IT systems become more agile and flexible. The external provision of cloud-based services, as part of cloud computing, appears as an evolution of traditional outsourcing, due to the emergence of emerging technologies related to IT service provision. As a result of technological development, traditional outsourcing and external provision of cloud-based services share common characteristics, such as cost reduction, increased flexibility, IT simplification and release of resources, among others. But there are some significant differences between the external provision of cloud-based services and traditional outsourcing, always from the point of view of the customer, such as shorter term contracts, more transparent costs, less project management, less IT government and less coordination, among others.

MM-2GES has been designed in order to serve as a public service and to fit all forms of current and future external service provision, within a mixed integration system of solutions. This way, the practical implementation of the model becomes flexible, agile and adaptable. Therefore, this research also adapts MM-2GES to external provision of cloud-based services, from the customer point of view, by doing several justified adjustments in the metrics table (see Table II), where every characteristic of the model is assessed. The adjustments have been done depending on the differences between traditional outsourcing and the cloud-based services provided externally.

The applicability of the model can be implemented in organizations that have deployed both IT service provision models, traditional outsourcing and cloud-based services provided externally, in order to achieve excellence in governance and management of all kind of IT services provided externally in organizations.

The model designed in this research shares a common feature there is in both models cloud-based services provided externally and traditional outsourcing. This feature is flexibility to be adapted to possible forms of service provision to come in organizations, due to the ongoing technological evolution. Therefore, MM-2GES can be adapted to any kind of external provision of services, but only after doing the necessary adjustments in the metrics table (see Table II).

Finally, on the basis of this research, by categorizing concepts and subconcepts with a specific focus on IT outsourcing, and designing an assessment tool along with the maturity model that allows independent and practical application of the model, and ultimately adapting the model to cloud-based services provided externally, this study seeks to allow higher education institutions under study to meet successfully the requirements of the the complex digital era of the internet.

References

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. and Zaharia, M., “A view of cloud computing,” Communications of the ACM, vol. 53(4), 2010.

[2] Clemons, Eric K. and Chen, Yuanyuan, “Making the decision to contract for cloud services: Managing the risk of an extreme form of IT outsourcing,” 44th Hawaii International Conference on System Sciences, 1-10, doi: 10.1109/HICSS.2011.292, 2011.

[3] Dhar, S., “From outsourcing to cloud computing: evolution of IT services,” Management Research Review, vol. 35(8), pp. 664-675, 2012.

[4] Fernández, Eugenio, “UNiTIL: Gobierno y gestión del TIC basado en ITIL,” III Congreso Interacadémico / itSMF España, 2008, last access on 5th Sept 2012 at http://www.uc3m.es/portal/page/portal/congresos_jornadas/congreso_itsmf/ UNiTIL%20Gobierno%20y%20Gestion%20de%20TIC%20basado%20en%20ITIL.pdf

[5] Gottschalk, P. and Solli-Saether, H., “Critical success factors from IT outsourcing theories: an empirical study,” Industrial Management & Data Systems, vol. 105(5/6), pp. 685-702, 2005.

[6] Joint, A. and Baker, E., “Knowing the past to understand the present issues in the contracting for cloud based services,” Computer Law & Security Review, vol. 27(4), pp. 407-415, 2011.

[7] Patel, Nandish, “Emergent forms of IT governance to support global eBusiness models,” Journal of Information Technology Theory and Application (JITTA), vol. 4 (2), article 5, 2002, last access on 19th July 2011 at http://aisel.aisnet.org/jitta/vol4/iss2/5.

[8] Rodder, N., Knapper, R., Martin, J., “Risk in modern IT service landscapes: Towards a dynamic model,” 5th IEEE International conference on Service-Oriented Computing and Applications (SOCA), 1-4, doi: 10.1109/SOCA.2012.6449445, 2012.

[9] Schwarz, C., “Toward an understanding of the nature and conceptualization of outsourcing success,” Information & Management, vol. 51(1), pp. 152-164, 2014.

[10] Thompson, S.H. Teo, Bhattacherjee, Anol, “Knowledge transfer and utilization in IT outsourcing partnerships: A preliminary model of antecedents and outcomes,” Information & Management, vol. 51(2), pp. 177-186, 2014.

[11] Weinhardt, C., Anandasivam, A., Blau, B., Borissov, N., Meinl, T., Michalk, W. and Stoesser, J., “Cloud computing – a classification, business models, and research directions,” Business & Information Systems Engineering, vol. 1(5), pp. 391-9, 2009.

Parallel Domain Decomposition for 1-D Active Thermal Control Problem with PVM

Simon Uzezi Ewedafe
Department of Computing
The University of the West Indies,
Mona Kingston 7, Jamaica

Rio Hirowati Shariffudin
Institute of Mathematical Sciences
Universiti Malaya
Kuala Lumpur, Nigeria

***Abstract*—this paper describes a 1-D Active Thermal Control Problem (1-D ATCP) with the use of Stationary Iterative Techniques (Jacobi and Gauss-Seidel) on the discretization of the resulted matrices. Parallelization of the problem is carried out using Domain Decomposition (DD) parallel communication approach with Parallel Virtual Machine (PVM) to enable better flexibility in parallel execution, and greater ease of parallel implementation across the different domain of block sizes. We described the parallelization of the method by the use of Single Program Multiple Data (SPMD) technique. The 1-D ATCP is implemented on a parallel cluster (Geo Cluster), with the ability to exploit inherent parallelism of the computation. The parallelization and performance strategies are discussed, and results of the parallel experiments are presented.**

Keywords—1-D ATCP; Stationary Techniques; SPMD; DD; PVM

I. Introduction

Researchers' efforts in parallel computing have focused on improving various aspects of communication performance in parallelism. The goal is to provide a study on the impact of communication performance for parallel applications. The focus is on high-performance cluster architecture, for which a fast active message-passing layer to a low latency, high bandwidth network is provided [20]. Applications are most sensitive to overhead, and some are hyper-sensitive to overhead in that the execution time increases with overhead at a faster rate that the message frequency would predict [2]. Developments in computer architectures have shown that parallel computer systems are a viable way to overcome major hardware design challenges relating to energy-consumption, and thermal constraints, while offering high computational peak performance [17]. Many applications are also sensitive to message to message rate, or to message transfer bandwidth, but the effect is less pronounced than with overhead. Cluster applications have more processor cores to manage, and exploit the computational capacity of high-end machines providing effective and efficient means of parallelism even as the challenges of providing effective resources management grows. It is a known fact that high capacity computing platform are expensive and are characterized by long-running, high processor-count jobs [3]. Developing parallel applications have its own challenges in the field of parallel computing. Regarding [12], there are theoretical challenges such as task decomposition, dependence analysis, and task scheduling.

Asynchronous message driven execution is a convenient and effective model for general purpose parallel programming. The flow of control in the message-driven systems is dictated by the arrival of messages. In Charm++ based message-driven applications [14], the problem to be solved is decomposed into collections of communicating parallel objects, providing the opportunity for easy overlap of communication with computation [7], and run time-level optimization such as automatic load balancing. In the loosely-coupled style, the assembly of separate parallel modules in a single application only requires interfaces, rather than the deeper structural changes or non-overlapped time, or processor division that might be required in the SPMD models [18]. In cases where data the concise expression of the algorithm, the code needed to explicitly communicate this data is a message-driven style, and this can dominate the structure of the program, and overwhelm the programmer as in [18]. For a global task with other processors relevant data needs to be passed from processor to a processor through a message-passing mechanism [19, 13], since there is greater demand for computational speed and the computations must be completed within reasonable time period. Design and analysis on finite difference DD for 2-D Heat Equations have been discussed in [24], and the parallelization for 3-DTEL on a parallel virtual machine with DD in [8] show effective load scheduling over various mesh sizes, which produce the expected inherent speedups.

This paper concern the estimated peak junction temperature of semiconductor devices during the manufacturing process, which is part of the thermal control systems [23] that treats the sequential algorithm of Parabolic Equation in solving thermal control process on printed circuit board. Temperature control of high-power microprocessor devices during testing is very important to determine the device performance [10]. The device manufacturer specifies a temperature and allowed deviation from it during the testing procedure (e.g., $85^{o}c$ – $o/+2^{o}c$). The test process applies computer-controlled electrical signals to the device, and for high-power devices, this may result in die-average power density in the range of $100KW/m^{3}$ [6]. High power microprocessor devices are also subjected to a classification test to determine the effective operating speed of the device. During this classification test, the goal of control is to keep the temperature of the die at a single set temperature where the device power is varied between 0% to 100% power in a predetermined test sequence [21]. An alternative approach to thermal management in automatic testing equipment was proposed and tested by [10]. In their approach, the surface of the device under test is heated with laser radiation while simultaneously cooled by forced convention. A significant complication in this scheme arises from the time required for temperature signal propagation from the device package surface to the die, upon which the power actually dissipated. However, [21] analyzed the effect of the conduction time lag on the control power for sinusoidal die power. Minimization of the required laser power, which can amount to hundreds of watts or more, is of great importance to limit both the electrical power consumed by the control system and the added load on the test facility's cooling system. Hence, [6] extended the analysis of [21] to multi-frequency wave forms, with the aid of determining optimal control power for multi-frequency test power sequences. They show that the profile control calculation with specified die temperature tolerance in [21] is not suitable for non-sinusoidal die power profiles, and they develop a new approach for the situation. All the high performance electronic devices are subjected to a 100% functional test prior to being shipped by the manufacturers [21].

The theoretical properties of the 1-D stationary techniques with the DD parallel communication approach with PVM to enable better flexibility in parallel execution, and greater ease of parallel implementation across the different domain of block sizes**,** and employing SPMD technique is emphasized in this paper. The 1-D ATCP is implemented on the Geo Cluster with the ability to exploit inherent parallelism. Previous works on 1-D stationary techniques on ATCP did not consider the parallel communication and DD approach on PVM to determine the temperature distribution. Our programming style allows the application programmer to specify the program organization in a clear and readable program code.

This paper presents a systematic methodology for investigating the effects of measuring variations in communication performance on the temperature distribution and in-depth study of application sensitivity to overhead and bandwidth in quantifying application performance in respond to changes in communication performance of the ATCP. In general, we use the above efforts to improve communication performance in Geo cluster architecture. Our results demonstrated flexibility in parallel execution, and greater ease of implementation. On the other hand, to improve the understanding of the stability of ATCP, this paper aims to confirm simulation to give increase confidence in the reality of the system. It also provides useful stationary techniques for evaluating thermal control problems even to higher dimensional problems.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 introduces the model for the 1-D ATCP and the stationary schemes. Section 4 introduces the parallel design and implementation details, including the PVM. Section 5 introduces the results of the experiments for the parallelization on Geo cluster. Section 6 gives the conclusion.

II. Related Work

The implementation of sequential algorithm in solving ATCP on printed circuit board with numerical finite difference method to design the discretization of the Partial Differential Equations (PDE) was implemented in [23]. The implementation design approaches are either passive or active controlled. The passive controlled design is suitable for low to medium power dissipation, while the active thermal control system is suitable for industrial processing, and testing of die. In testing package high-power integrated circuits, active thermal control is useful in providing die-level temperature stability [6]. They consider active test power sequences that contain many frequencies at various phase angles, each contributing to the temperature of the die. They develop a method of temperature variation of the die, and a method of controlling multiple frequency test sequences subject to a finite temperature tolerance. Sweetland et al. [21] presented an active thermal control of distributed-parameter system; with application to testing of packaged integrated circuit devices, requiring the die-level temperature be modulated to account for the distributed thermal capacitance and resistance of the device packaging. They also demonstrated fundamental limits of temperature control for typical devices under test conditions. Parallelization by time decomposition was first proposed by [16] with motivation to achieve parallel real-time solutions, and even the importance of loop parallelism, loop scheduling have been extensively studied [1]. Programming on heterogeneous many-core systems using explicit platform description to support programming by [17] constituted a viable approach for coping with power constraints in modern computer architectures, and can be formed across the whole computing landscape to high-end supercomputers and large-scale data centers. The stationary iterative techniques have been developed to solve various problems in PDEs [4], been widely used for solving algebraic systems resulting from the use of finite difference methods in several scientific and engineering applications. Chi-chung et al. [5] used a network of workstations as a single unit for speeding up computationally intensive applications as a cost-effective alternative to traditional parallel computers. The platform provided a general and efficient parallel solution for time-dependent PDE. However, [22] proposed an efficient parallel finite-difference scheme for solving Heat Equations numerically. They based it upon the overlapping DD method. Ewedafe and Rio [8] parallelized a 3-D ADI scheme using DD method with the SPMD technique on a MPI platform of clusters. On the same investigation, [9] used a numerical iterative scheme to solve 2-D Bio-Heat on MPI/PVM cluster systems with the SPMD technique. The Geo cluster are designed for application running on distributed memory clusters which can dynamically and statically calculate partition sizes based on the run-time performance of applications. We use the stationary iterative techniques (Jacobi and Gauss-Seidel) on the resulted matrices from the discretization of the 1-D ATCP model. Parallelization of the problem is carried out using DD parallel communication approach with PVM. The parallelization strategy and performance are discussed, and results of the parallel experiments are presented.

III. The Model Problem

The mathematical model for the 1-D ATCP follows the 1-D model of [21]. For the transient response, the physical model of the device is reduced to 1-D model of the form:

$$\frac{\partial^2 v(x,t)}{\partial x^2} = \frac{1}{(b_t)}\frac{\partial v(x,t)}{\partial t} \tag{3.1}$$

where b_t is the thermal diffusivity (that measure the ability of a material to conduct thermal energy relative to its ability to store thermal energy), and $b_t = \alpha = \dfrac{k}{\rho c_p}$,

and k is the thermal conductivity $(W/(m.k))$, ρ is the density (kg/m^3), and c_ρ is the specific heat capacity $(J/(kgk))$ while ρc_p together can be considered the volumetric heat capacity $(J/(m^3k))$. Hence,

$$\begin{gathered} b_t \frac{\partial^2 v(x,t)}{\partial x^2} = \frac{\partial v(x,t)}{\partial t} \\ \text{let } v(x,t) = V, \text{then we have} \\ b_t \frac{\partial V}{\partial x^2} = \frac{\partial V}{\partial t} \end{gathered} \tag{3.2}$$

We can then solve (3.2) by extending the 1-D explicit finite difference method to the above, Eq. (3.2) becomes:

$$\begin{aligned} \frac{\partial V}{\partial t} &= V_t = \frac{V_{i,j+1} - V_{i,j}}{\Delta t} \\ \frac{\partial^2 V}{\partial x^2} &= V_{xx} = \frac{V_{i+1,j} - 2V_{i,j} + V_{i-1,j}}{(\Delta x)^2} \end{aligned} \tag{3.3}$$

applying eq. (3.3) on Eq. (3.2), the temperature of the explicit node is given by:

$$\frac{V_{i,j+1} - V_{i,j}}{\Delta t} = b_t \left(\frac{V_{i+1,j} - 2V_{i,j} + V_{i-1,j}}{(\Delta x)^2} \right)$$

rearranging we have:

$$V_{i,j+1} = \frac{bt\Delta t}{(\Delta x)^2}\left(V_{i+1,j} - V_{i-1,j}\right) + \left(1 - 2b_t \frac{\Delta t}{(\Delta x)^2}\right)V_{i,j} \tag{3.4}$$

although an implicit scheme of the Crank-Nicolson (C-N) unconditionally stable scheme cab be applied to Eq. (3.2) as seen below.

A. C-N Implicit Scheme on 1-D ATCP

When the C-N implicit scheme is used we write using the same finite difference scheme:

$$\frac{V_{i,j+1} - V_{i,j}}{\Delta t} = \frac{b_t}{2}\left(\frac{\begin{gathered} V_{i+1,j+1} - 2V_{i,j+1} + V_{i-1,j+1} \\ + V_{i+1,j} - 2V_{i,j} + V_{i-1,j} \end{gathered}}{(\Delta x)^2} \right) \tag{3.5}$$

and rearranging to give:

$$\begin{gathered} -b_t \frac{\Delta t}{(\Delta x)^2} V_{i-1,j+1} + \left(2 + b_t \frac{\Delta t}{(\Delta x)^2}.2 \right) V_{i,j+1} \\ -b_t \frac{\Delta t}{(\Delta x)^2} V_{i+1,j+1} = b_t \frac{\Delta t}{(\Delta x)^2} V_{i-1,j} + \\ \left(2 - b_t \frac{\Delta t}{(\Delta x)^2}.2 \right) V_{i,j} + b_t \frac{\Delta t}{(\Delta x)^2} V_{i+1,j} \end{gathered} \tag{3.6}$$

let $b_t \dfrac{\Delta t}{(\Delta x)^2} = r$, then we have:

$$\begin{gathered} -rV_{i-1,j+1} + (2+2r)V_{i,j+1} - rV_{i+1,j+1} = \\ rV_{i-1,j} + (2-2r)V_{i,j} + rV_{i+1,j} \end{gathered} \tag{3.7}$$

here, we have a tridiagonal system of equations which can be solved with the stationary iterative techniques.

B. Stationary Techniques on 1-D ATCP

A system of linear algebraic equations can be sparse and banded. We will typically employ the concise notation

$$Au = b \tag{3.8}$$

to represent such systems and the focus of this section is the study of methods for efficiently solving equation (3.8) on parallel computers. We begin with the decomposition

$$A = D - E - F, \tag{3.9}$$

in which D is the diagonal of A, $-E$ is the strict lower part and $-F$ is the strict upper part. It is always assumed that the diagonal entries of A are all nonzero. The Jacobi iteration determines the ith component of the next approximation so as to annihilate the ith component of the residual vector.

Thus,

$$(b - Ax_{k+1}) = 0 \tag{3.10}$$

however, recall Eq. (3.8) and note that iterative methods for solving this system of linear equations can essentially always be expressed in the form:

$$U^{(n+1)} = GU^{(n)} + K \tag{3.11}$$

where n is an iterative counter and G is the iteration matrix, it is related to the system matrix A by

$$G = I - Q^{-1}A$$

where I is the identity matrix and Q is generally called the splitting matrix. The Jacobi scheme can be constructed as follows. Firstly, decompose A as in Eq. (3.9), substitute into (3.8) to obtain:

$$(D-L-U)U=b,\ or\ DU=(L+U)U+b \qquad (3.12)$$

hence, introducing iteration counter, (3.12) becomes

$$U^{(n+1)}=D^{-1}(L+U)U^{n}+D^{-1}b \qquad (3.13)$$

from Eq. (3.13) $L+U=D-A$, so
$D^{-1}(L+U)=I-D^{-1}A.$

Thus, D is the splitting matrix and Eq. (3.13) is in the form (3.11) with

$$G\equiv D^{-1}(L+U)=1-D^{-1}A,\ k=D^{-1}b \qquad (3.14)$$

hence, in matrix terms the definition of the Jacobi method can be expressed as

$X^{(k+1)}=D^{-1}(L+U)x^{(k)}+D^{-1}b$ as in Eq.(3.13)

where

$$x_i=\frac{1}{a_{i,j}}(b_i-\sum_{j\neq i}a_{i,j}x_j) \qquad (3.15)$$

suggesting an iterative method defined by

$$x_i^{(k+1)}=\frac{1}{a_{i,j}}(b_i-\sum_{j\neq i}a_{i,j}x_j^{k}) \qquad (3.16)$$

Consider again the linear equations; if we proceed as with the Jacobi method but not assume that the equations are examined one at a time in sequence, and that previously computed results are used as soon as they are available we obtain the Gauss-Seidel (GS):

$$X_i^{(k)}=\frac{1}{a_{i,j}}\left[b_i-\sum_{j<i}a_{i,j}x_j^{(k)}-\sum_{j>i}a_{i,j}x_j^{(k-1)}\right] \qquad (3.17)$$

the computation appear to be serial, since each component of the new iterate depends upon all previously computed components. The update cannot be done simultaneously as in the Jacobi method. Secondly, the new iterate $x^{(k)}$ depends upon the order in which the equations are examined. The GS is called the "Successive Displacement" to indicate the dependence of iterates on the ordering. A poor choice of ordering can degrade the rate of convergence.

IV. Parallel Implementation, Design and Analysis

A. The Parallel Platform

The implementation was done on Geo Cluster consisting of 16 Intel Celeron CPU J 1900 at 1.99GHz quad core, and 495.7GB of Disk type. PVM [11] is a software system that enables a collection of heterogeneous computers to be used as a coherent and flexible concurrent computational resource. It supports program executed on each machine in a user-configurable pool, and present a unified, general, and powerful computational environment for concurrent applications. The program, written in Fortran, C, or C++, are provided access to PVM through calling PVM library routines for functions such as process initiation, message transmission and reception, and synchronization via barriers or rendezvous. PVM is ideally suited for concurrent applications composed o many interrelated parts, and is very useful for the study of large-scale parallel computation.

B. Parallel Implementation and DD

Partitioning strategy simply divides the problem into parts. Most partitioning formulations, however, require the results of the parts to be combined to obtain the desired result. Partitioning can be applied to the program data i.e. dividing the data and operating upon the divided data concurrently. This is called data partitioning or DD. Decomposition into fixed chunks per processing element. When the domain is splinted, each block is given an identification number by a "master" task, which assigns these sub-domains to "slave" tasks running in individual processors. In order to couple the sub-domains' calculations, the boundary data of neighboring blocks have to be interchanged after each iteration. The calculations in the sub-domains use the old values at the sub-domains' boundaries as boundary conditions. DD is used to distribute data between different processors; the static load balancing is used to maintain same computational points for each processor. Data parallelism originated the SPMD [15], thus, the finite difference approximation used in this paper can be treated as an SPMD problem. The SPMD model contains only a single program with multiple data and each process will execute the same code. To facilitate this within a single program, statements need to be inserted to select which portions of the code will be executed by each processor. The copy of the program is started by checking pvm_parent, it then spawns multiple copies of itself and passes then the array of tids. At this point, each copy is equal and can work on its partition of the data in collaboration with other processes. In the master model, the master program spawns and direct a number of slave program which perform computations. Any pvm task can initiate processes on the machine. The master calls pvm_mytid, which as the first pvm call, enrolls this task in the pvm system. It then calls pvm_spawn to execute a given number of slave programs on other machines in pvm. Each slave calls pvm_tid to determine its task id in the virtual machine, and then uses the data broadcast from the master to create a unique ordering from 0 to nproc minus 1. Subsequently, pvm_send and pvm_recv are used to pass messages between processors. When finished, all pvm programs call pvm_exit to allow pvm to disconnect any sockets to the process and keep track of which processes are currently running. A master program wakes up worker programs, assign initial data to the workers and let them work, receive results from workers, update and display them. The worker program works like this: receive initial data from master, exchange the edges data with the next door workers, and send the result to the master.

C. Parallel Communication with PVM

The activities of communication are slow compare to computation, often by order of magnitude. The communication costs are measured in terms of latency and bandwidth. The communication costs comprised of the physical transmission costs and the protocol overhead. The overhead is high in heterogeneous networks, where data may have to be converted to another format. The communication cost can be reduced but not avoided. The non-blocking communication is employed across SPMD to reduce the problem of blocking. Factors considered in communication include; Cost of communications which implies overhead and required synchronization between tasks, latency versus bandwidth, synchronization communication, and efficiency of communication. We have our grid distributed in a block fashion across the processors, the values for the ghost cells are calculated on neighboring processors and sent using PVM class. Synchronization is the coordination of parallel tasks associated with communication. A routine that returns when the transfer has been completed is a synchronous message passing and performs two actions: transferring of data then synchronous processes through send / receive operations

D. Speedup and Efficiency with Effectiveness

Speed-up and efficiency are commonly used to measure the performance of a parallel code. The runtime of the original serial code is used as a measured of the runtime on one processor. In this context, runtime can be defined as the time that has elapsed from the first moment when the first processor actually begins execution of the program to the moment when the last process executes its last statement. In this present code, time is measured after the initialization of PVM, and before the domain decomposition. The time measurement ends after the writing of the last result, just before finalizing PVM. Only the timing of the day is considered. The parallel Speed-up (Sp) is the ratio of the runtime on one processor t1 to the runtime on P processor tp.

$$Sp = \frac{t1}{tp} \tag{4.1}$$

The parallel efficiency Ep is the ratio of the parallel Speed-up Sp to the number of processors P.

$$Ep = \frac{Sp}{p}, \tag{4.2}$$

another intuitive aspect for the performance but also a value easy to obtain for optimizing the efficiency of the parallel code is the communication time.

Hence, effectiveness is given as

$$L_n = S_n / (nT_n) = E_n / T_n = E_n S_n / T_1 \tag{4.3}$$

which clearly shows that L_n is a measure of both speedup and efficiency. Therefore, a parallel algorithm is said to be effective if it maximizes L_n and hence $L_n T_1 = S_n E_n$.

V. Results and Discussion

Consider the Telegraph Equation of the form:

$$\frac{\partial^2 v(x,t)}{\partial x^2} = \frac{1}{(b_t)} \frac{\partial v(x,t)}{\partial t} \tag{5.1}$$

the boundary condition and initial condition posed are:

$$\begin{aligned} &V(0,t) = h(t), \\ &V(1,t) = k(t), \qquad 0 \le t \le T \\ &\frac{\partial V}{\partial x}(x,0) = f(t), \quad 0 \le t \le 1 \\ &V(x,0) = f(x) \ (0 \le x \le 1) \end{aligned} \tag{5.2}$$

A. Parallel Efficiency

The speedup and efficiency obtained for various sizes, for 100 mesh size to 300 mesh size, are listed in Tables I to III, and the effectiveness of the model is shown in Table IV. The Tables show the parallel time decreasing as the number of processors increase for using the stationary iterating schemes. The speedup and efficiency versus the number of processors are shown in Fig. 1 and Fig. 2, respectively. The results in the Tables show that the parallel efficiency increases with increasing grid size, and decreases with the increasing block number for given grid size. As the number of processors increase, though this leads to a decrease in execution time, but a point is reached when the increased processors will not have much impact on total execution time. Hence, when the number of processor increase, balancing the number of computational cells per processors will become a difficult task due to significant load imbalance. Performance begins to degrade with an effect caused by the increase in communication overhead as the mesh increases. The gain in increasing execution time for certain mess sizes is due to uneven distribution of the computational cell, and the execution time has a very small change due to DD influence on performance in parallel computation.

TABLE I. The Total Time T, The Parallel Speed-Up Spar and the Efficiency Epar for a Mesh of 100

Scheme	*N*	*S_{par}*	*E_{par}*
Jacobi	*1*	*1.000*	*1.000*
	2	*0.953*	*0.477*
	6	*1.048*	*0.175*
	8	*1.296*	*0.162*
	10	*1.564*	*0.156*
	12	*1.831*	*0.153*
	14	*2.094*	*0.150*
	16	*2.313*	*0.145*
GS	*1*	*1.000*	*1.000*
	2	*1.108*	*0.554*
	6	*1.216*	*0.203*
	8	*1.397*	*0.175*
	10	*1.645*	*0.165*
	12	*2.109*	*0.176*
	14	*2.271*	*0.162*
	16	*2.526*	*0.158*

TABLE II. The Total Time *T*, the Parallel Speed-Up S_{PAR} and the Efficiency E_{PAR} for a Mesh of 200

Scheme	*N*	*S_{par}*	*E_{par}*
Jacobi	*1*	*1.000*	*1.000*
	2	*1.074*	*0.537*
	6	*1.265*	*0.211*
	8	*1.532*	*0.192*
	10	*1.848*	*0.185*
	12	*2.226*	*0.186*

	14	*2.584*	*0.185*
	16	*2.928*	*0.183*
	1	*1.000*	*1.000*
	2	*1.205*	*0.603*
	6	*1.328*	*0.221*
GS	*8*	*1.612*	*0.202*
	10	*1.924*	*0.192*
	12	*2.497*	*0.208*
	14	*2.703*	*0.193*
	16	*3.034*	*0.190*

TABLE III. THE TOTAL TIME *T*, THE PARALLEL SPEED-UP S_{PAR} AND THE EFFICIENCY E_{PAR} FOR A MESH OF 300

Scheme	*N*	S_{par}	E_{par}
	1	*1.000*	*1.000*
	2	*1.348*	*0.674*
	6	*1.665*	*0.278*
Jacobi	*8*	*2.038*	*0.255*
	10	*2.606*	*0.261*
	12	*3.115*	*0.260*
	14	*3.665*	*0.262*
	16	*4.190*	*0.262*
	1	*1.000*	*1.000*
	2	*1.441*	*0.721*
	6	*1.813*	*0.302*
GS	*8*	*2.176*	*0.272*
	10	*2.747*	*0.275*
	12	*3.419*	*0.285*
	14	*3.857*	*0.276*
	16	*4.317*	*0.270*

TABLE IV. EFFECTIVENESS OF THE VARIOUS SCHEMES WITH PVM FOR 300 MESH SIZE

	N	*PVM T(s)*	L_n
	1	*178.5*	*0.56*
	2	*132.4*	*0.51*
Jacobi	*6*	*107.2*	*0.26*
	8	*87.6*	*0.29*
	10	*68.5*	*0.38*
	12	*57.3*	*0.45*
	14	*48.7*	*0.54*
	16	*42.6*	*0.55*
	1	*213.7*	*0.47*
GS	*2*	*148.3*	*0.46*
	6	*117.9*	*0.26*
	8	*98.2*	*0.28*
	10	*77.8*	*0.35*
	12	*62.5*	*0.46*
	14	*55.4*	*0.50*
	16	*49.5*	*0.55*

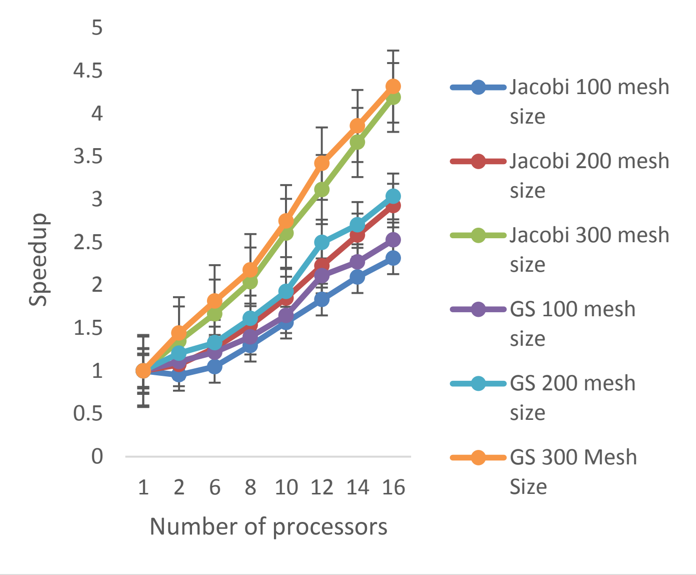


Fig. 1. Speedup versus the number of processors for mesh 100,200 and 300

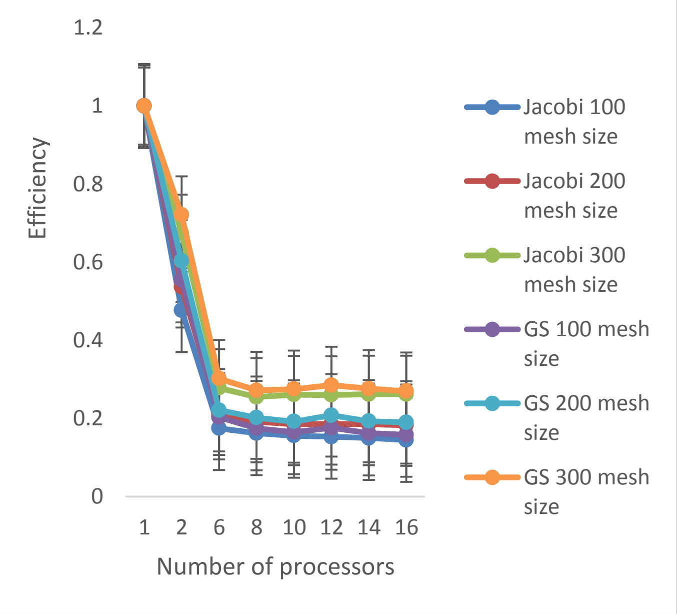


Fig. 2. Parallel efficiency versus the number of processors for mesh 100, 200 and 300

The numerical efficiency includes the DD efficiency and convergence rate behavior. The DD efficiency includes the increase of floating point operations induced by grid overlap at interfaces and the CPU time variation generated by DD techniques. , the performance is strongly dependent on the load condition of the nodes, and the distributed platform allows efficient overlapping of computation and communication. It is pointed out that when the computational domain are partitioned over the processors, especially for small domains, the execution time and speed-up seems not noticeable, but with larger mesh sizes the effect of parallelization become noticeable. The result presented in the Tables and Figures show that PVM is both powerful and economically distributed processing tool.

VI. Conclusion

In this paper, we consider the combination of the stationary iterative techniques on the resulted matrices from the discretization process of the 1-D ATCP model with the SPMD parallel domain decomposition technique, which sacrifices some flexibility of a parallel platform with the general-purpose message-driven execution. To demonstrate the use of the stationary techniques on the above-mentioned model, we present example of the model with initial and boundary conditions implemented on different mesh sizes. The temperature control in a distributed parameter thermal system has been extended to parallel implementation on cluster systems in one dimension. The results presented in this paper show the study of the parallel domain decomposition on a parallel cluster and analysis for the above model with PVM. The system allows a parallel collection of overlapping communication. Computational results obtained have clearly shown the benefits of parallelization. The DD greatly influences the performance of the 1-D ATCP model on the parallel clusters. However, we are interested in improving the algorithms used in this paper, and also to test the improved algorithms on increase number of processors thereby exploits the communication of data transfer across multiple clusters.

VII. Future Work

The parallel domain decomposition for 1-DATCP with the use of Stationary Iterative Techniques on Geo Cluster Systems using PVM employing the SPMD technique has been carried out. This paper has shown the ability to exploit inherent parallelism of the computation. We suggest future work to be carried out on the 1-D ATCP model employing the used of Iterative Alternating Direction Implicit (IADE) method. Parallel implementation on the scheme could use the Input File Affinity Measure on a tightly coupled and loosely coupled distributed environment with dynamic allocation of a task with varying mesh sizes. Another avenue for improvement would be to make the parallel implementation adapt a unique pattern of predictable based knowledge of the algorithms on problem domain in question.

Acknowledgments

The University of the West Indies Mona Campus awarded the Initiative Grant that motivated this study.

References

[1] J. Aguilar, E. Leiss, ‘Parallel Loop Scheduling Approaches for Distributed and Shared Memory System’, Parallel Process Letter 15 (1 – 2), 2005, pp. 131 – 152

[2] E. Aubanel, ‘Scheduling of tasks in the parareal algorithm’ Parallel Computing 37 (3), 2011, 172 – 182

[3] W. Barry, A. Michael, ‘*Parallel Programming Techniques and Application using Networked Workstation and Parallel Computers*’ 2003, Prentice Hall, New Jersy

[4] R. L. Burden, J. D. Faires ‘Numerical Analysis 5th ed.’ PWS Publishing Company, Boston, 1993

[5] H. Chi-chung, Ka-keung C., G. Man-Sheung Yuen, M. Hamdi, Solving PDE on Network of Workstation. IEEE, 1994, pp194 – 200

[6] C. R. Christopher, J. H. Lienhard, ‘Active Thermal Control of Distributed Parameter Systems Excited at Multiple Frequencies’ Journal of Heat Transfer, 2005

[7] S. U. Ewedafe, H. S. Rio, ‘Armadillo Generation Distributed Systems & Geranium Cadcam Cluster for solving 2-D Telegraph Equation’ Int’l Jour. of Computer Mathematics, 88, Issue 3, 2011, 589 – 609

[8] S. U. Ewedafe, H. S. Rio, ‘On the Parallel Design and Analysis for 3-D ADI Telegraph Problem with MPI’ Int’l Jour. Of Advanced Compt. Sci. and Applications, Vol. 5, No. 4, 2014

[9] S. U. Ewedafe, H. S. Rio, ‘Domain Decomposition of 2-D Bio-Heat Equation on MPI/PVM Clusters’ Int’l Journal of Emerging Trends of Technology in Compt. Science, Vol. 3, Issue 5, 2014

[10] P. Fahnl, A. C., Lienhard V., J. H, A. H. Slocum ‘Thermal management and Control in Testing Packaged Integrated Circuit (IC) Devices’ Proc. 34th Inter Society Energy Conversion Conference’ 1999

[11] A. Geist, A. Beguelin, J. Dongarra, ‘Parallel Virtual Machine (PVM)’ Cambridge MIT Press

[12] N. Giacaman, O. Sinnen, ‘Parallel iterator for parallelizing object-oriented applications’ Intl journal of parallel programming, 39 (2), 2011, 223 – 269, 2011.

[13] K. Jaris, D.G. Alan, ‘A High-Performance Communication Service for Parallel Computing on Distributed System’, Parallel Computing 29, 2003, pp 851 – 878

[14] L. V. Kale, S. Krishnan, ‘Charm++ Parallel Programming with Message-Driven Objectives, in: G. V. Wilson, P. Lu (Ed.)’ Parallel Programming using C++, MIT Press, 1966, pp 175 - 213

[15] H. Laurant, ‘A method for automatic placement of communications in SPMD parallelization’ Parallel computing 27, 2001, 1655 – 1664

[16] J. L. Lions., Y. Maday, G. Turinki, ‘Parareal in time discretization of PDE’ Comptes, rendus de lacadimie des sciences – series 1 – mathematics 332 (7), 2011, 661 – 668

[17] S. Martin, S. Benkner, S. Pllans, ‘Using Explicit Parallel Description to Support Programming of Heterogeneous Many-Core Systems’, Parallel Computing 38 (2012), pp 52 – 65

[18] P. Miller, A. Becker, L. Kale, ‘Using Shared Arrays in Message-Driven Parallel Programs’ Parallel Computing 38 (2012), pp 66 - 74

[19] Peizong L., Z. Kedem, ‘Automatic Data and Computation Decomposition on Distributed Memory Parallel Computers’ ACM Transactions on Programming Languages and Systems, vol. 24, number 1, 2002, pp 1 – 50

[20] P. Richard, M. Amin, E. David, E. Thomas, *‘The work of Parallelism*’ Computer Science University of California Berkeley CA 94720

[21] M. Sweetland, V. J. H. Lienhard, ‘Active Thermal Control of Distributed Parameter System with Application to Testing of Packaged (IC) Devices’ ASME Journal of Heat Transfer, 2003, pp 165 – 174

[22] M. Tian, D. Yang, ‘Parallel Finite Difference Schemes for Heat Equations Based on Overlapping Domain Decomposition”, Applied Maths & Compt, Issue 18, 2007, pp 1276 – 1292

[23] S. Zarith, A. Ghaffar, N. Alias, F. Sham, A. Hassan, H. Hassan, ‘Sequential Algorithms of Parabolic Equation in Solving Thermal Control Process on Printed Circuit Board’ Jour. Fundamental Science Issue 4, 2008, pp 379 - 385

[24] W. Zheng-Su, Z. Baolin, C. Guang-Nan, ‘Design and analysis for finite difference DD for 2-D heat equation’, ICA3PP – 02, IEEE Computer Society

Lempel - Ziv Implementation for a Compression System Model with Sliding Window Buffer

Ahmad AbdulQadir AlRababah
Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University,
Rabigh 21911, Kingdom of Saudi Arabia

***Abstract*—Proposed compression system architecture based on Lempel-Ziv algorithm with sliding window history buffer, this architecture may be realized on FPGA, and can provide input data streams from multiple sources and context switching. Base requirements to impression system and compression system architecture were proposed. Compression system architecture should provide quick reconstruction possibility for building another system with other technical characteristics and architecture features (like reconfigurable system architecture features) on given architecture base. Digital signal processing may comprise lined or non-lined procedures. Non-lined signal processing is strictly associated with no lined structure sympathy then can apply in this period, regularity, and patio-temporal fields.**

***Keywords*—Digital signal processing; FPGA; RAM; dual-port RAM; token; literal**

I. Introduction

Information technology penetration into various areas of life and industry provokes avalanche-like increasing information sites, which should be processed, saved and transported [1,4]. In addition to this increasing user's number of metropolitan, and wide area networks causes necessity of economical and more effective usage of data transmission for providing quality service, to all users [2,6,7]. While data processing algorithms, which perform some functions, are application specific, data transmission and saving algorithms must be universal for providing the ability to use them in various areas [5,12,14].

Compression systems used for more effective using of saving and transmission system resources [3]. Lossless compression systems, which provide restoring compressed information without distortion, should be distinguished. Arithmetic encoding algorithms various modifications of Huffman encoding algorithms and Lempel-Ziv compression algorithms are examples of Universal Lossless compression algorithms [8,11]. Basic demands to compression systems, compressor and decompressor architecture, based on Lempel-Ziv algorithms with sliding window history buffer, are considered in this work [12,14].

II. Basic Demands to Compression Systems

The main task of compression systems is to decrease information volume, which should be transferred or saved, typical demands are used for most of them, although, various application areas can require they won specific demands. We can distinguish such basic demands to compression systems:

- Providing high compression ratio for various data types. This parameter is algorithm dependent.
- Minimization of information volume was increasing (efficiency loss) when an unsuccessful condition for giving algorithm is met. Some compression algorithms can increase information volume during processing input data without or with low redundancy. Efficiency loss minimizations possibility is dependent from given algorithm and compression system architecture. For example, we can use input data stream analyzer, which can switch to another compression algorithm or turn off compression (bypass input stream to output without compression), when efficiency loss take place on given data stream.
- Providing high speed of information compression. Implementation of this demand depends on the compression algorithm and selected realization methods. For example, some algorithms perform preceding analysis before the start of compression process (two-pass algorithms). Data compression speed depends of selected realization methods (such as ASIC, FPGA, DSP, universal normal CPU, etc.), clock rate and other device specific parameters too.
- Providing possibility to process several independent input stream (multi-channel devices) with high-speed switching between them (context switching). This demand is very useful for network systems, where multiple input transmission channels multiplexed into one or several output channels.
- Providing a suitable interface for connecting compression systems with other systems.
- Universality providing. Compression system architecture should designed for using in different application areas without changing it for each new application.

III. Compressor Architecture

Compression processor architecture, based on random access memory proposed in this chapter. Architecture, based on register memory, when history buffer realized as big shift register with comparison in all cells at a time is more effective but needs to design new chip with complex signalization scheme from each cell. This architecture cannot be realized on FPGA due significant hardware requirements.

IV. Field Programmable Gate Array (FPGA)

FPGAs are programmable semiconductor campaigns that are established about an environment of Configurable Logic Blocks (CLBs) linked over programmable intersects. As per contrasting to Application Specific Integrated Circuits (ASICs), wherever the method is habit constructed aimed at the precise strategy, FPGAs can be automated to the favorite solicitation or functionality necessities. While One-Time Programmable (OTP) FPGAs are vacant, the governing kind are SRAM-based which can be reprogrammed as the project progresses.

FPGAs permit designers to modification their projects exact late in the project phase– unfluctuating subsequently the completion product has been mass-produced and installed in the field. Totaling, Xilinx FPGAs tolerate for field improvements to remain finalized tenuously, rejecting the charges allied with re-designing or manually apprising automatic methods.

An **ASIC** (application-specific integrated circuit) is a chip intended for a superior solicitation, such equally a specific type of diffusion procedure or a hand-held processer. You might compare it with universal combined circuits, such as the microchip and the random access memory chips in your PC. ASICs are used in a wide-range of submissions, as well as auto emanation controller, conservational monitoring, and personal digital assistants (PDAs). An ASIC may be pre-contrived for a distinct solicitation or it can be norm contrived (classically by modules from a "building block" archive of modules) for a specific client presentation.

Digital signal processing (DSP)

Digital signal processing (DSP) is the calculated employment of an evidence signal to transform or progress it in certain tactic. It is categorized by the depiction of detached period, separate occurrence, or other distinct dominion indications by a series of records or ciphers and the handling of these indications.

The aim of DSP is typically to ration, sieve and/or poultice incessant actual similarity signs. Commonly, the first stage is translation of the indication from an analog to a digital form, by selection and then digitizing it by an analog-to-digital converter (ADC), which cracks the analog signal into a torrent of separate numerical standards. Frequently, still, the requisite production sign is likewise analog, which necessitates a digital-to-analog converter (DAC). Smooth if this procedure is extra composite than analog handling and has a detached rate choice, the solicitation of computational control to sign handling consents for various benefits completed analog handling in numerous presentations, such as fault discovery and modification in communication as fine as data compression.

Digital and analog signal processing are subparts of signal processing. DSP submissions embrace audial and dialog sign handling, sonar and detector signal processing, sensor collection processing, phantom approximation, arithmetical sign processing, cardinal image processing, signal handling for infrastructures, controller of systems, amongst others. DSP procedures have elongated stayed route on regular processers, as well as on specific computers named alphanumeric signal processors, and on purpose-constructed hardware such as application-specific integrated circuit (ASICs). Presently, nearby are supplementary machineries used for digital signal processing counting extra controlling universal determination computer chip, field-programmable gate arrays (FPGAs), alphanumeric gesture regulators , and tributary mainframes, among others.

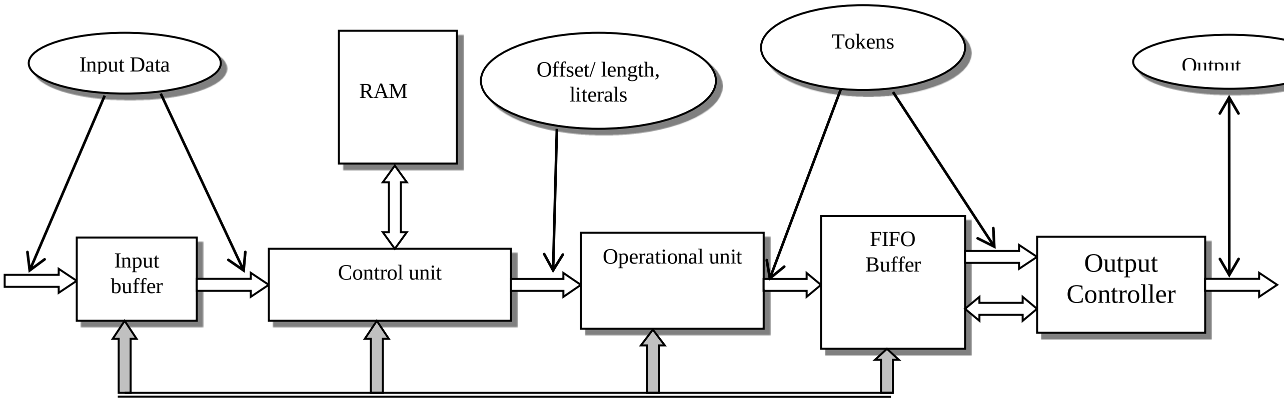


Fig. 1. Compressor Architecture

Compressor consists of the following units:

- Input buffer;
- RAM;
- Control unit;
- FIFO buffer;
- Output controller.

Input buffer used for buffering input data stream. Control unit intends to input data stream analysis, redundancy searching in this stream, token generation (literals and offset/length pair) and controlling other compressor units. Control unit is a main block in the system and consider as a more complex unit. Compressor productivity depends of efficiency of control unit and RAM cooperation. Operational unit used for literals and offset/length pair's modification and token generation according to algorithm specification. Operational unit controlled by a control unit. FIFO buffer devoted for tokes accumulation when output controller cannot produce output stream on the fly or when external device-destination of compressed information, is slower than compressor. Output controller used for forming compressed output data stream from tokens sequence. Output data stream may be either parallel or serial, so output controller form required output stream from tokens with variable length.

The main advantage of this architecture is simplicity and commonality of used component parts, what allow to build system on already created and tested components, reduce cost and time designing process. In addition, this architecture can be realized on FPGA, can provide input data streams form multiple sources and context switching, although, it needs additional memory for all possible sources and complex control unit.

One more of the main disadvantages of this architecture is significant duration of compressing process. This caused by consecutive nature of buffer memory, what provokes necessity of scanning whole buffer, for performing comparison, for each data element from input stream. It needed to complicate control unit and using, additional memory for storing intermediate comparison results.

V. DECOMPRESSOR ARCHITETURE

Decompressor consists of the following units:

- Input controller
- Operational unit
- Token buffer
- Main controller
- Circular buffer
- Dual-port RAM
- Control logic

Input controller is used for passing compressed input data stream and separate it into tokens. Tokens, in this algorithm, are independent parts of compressed information, which can belong to one of the two types: literal or offset/length token pair. Literal is modified (modification is algorithm dependent) part of information, which was not compressed. Offset/length pair is compressed part of information where offset is offset from start of history buffer where replication begins and length-length of replication part. Operational unit is used for tokens restoring from modifications, which can took place during compression process. Token buffer is used for storing tokens when output stream is much bigger than input stream, what can occur during processing compressed information with high compression ratio. Main controller is devoted for controlling all other units. Main controller takes tokens from token buffer and sends them to circular buffer, where, dependent from token type, some operations are fulfilled. Circular buffer (history buffer) is used for storing previously decompressed information and using it in next stages of decompression process. Circular buffer consists of Dual-port RAM unit and Control logic unit. Control logic is used for building circular buffer from Dual-port RAM and for tokens processing.

Proposed architecture is flexible and can be adapted for different variants of decompression algorithm. Most of changes will be applied to Input and Operational unit. This architecture can be used in multi-channel mode with context switch ing. Decompressor needs to have Token buffer and Dual-port RAM unit for each stream to achieve this,goal.

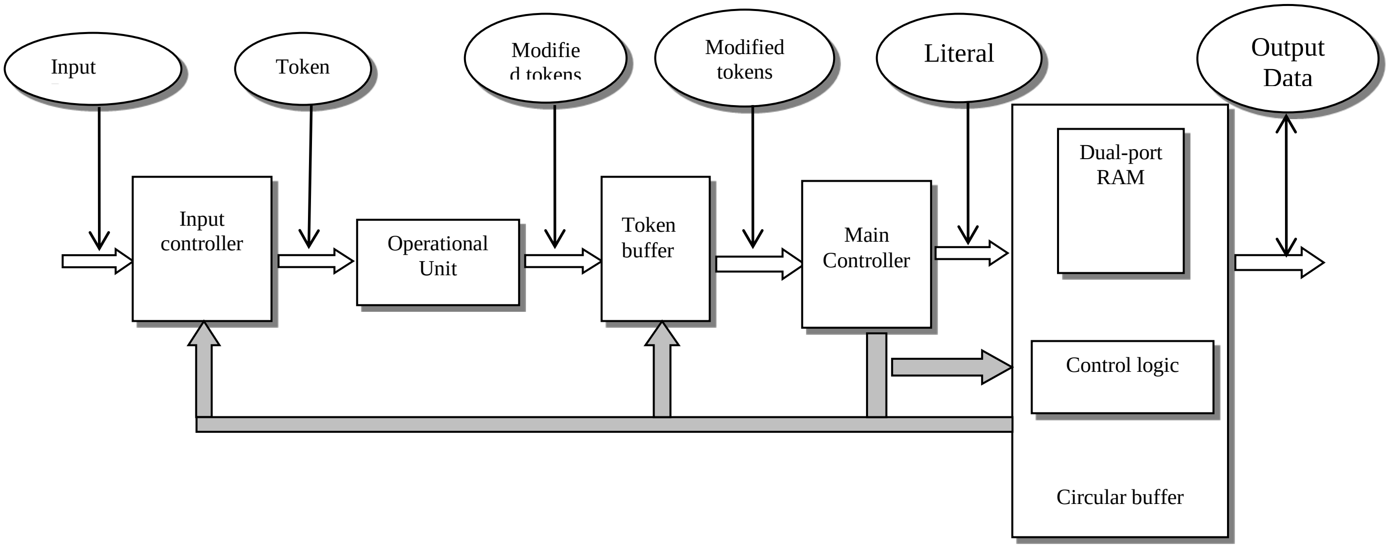


Fig. 2. Decompressor Architecture

V. Conclusion

Proposed compression system architecture based on Lempel-Ziv algorithm with sliding window history buffer, corresponds for more demands to compression systems and is an efficient solution for modern digital systems. One of the most important advantages of proposed architecture is the possibility to realize it in FPGA, that simplifies testing, reduces designing time and time for putting into operation.

References

[1] Khashman A., Dimililer K., (2008)," Image Compression Using Neural Networks And Haar Wavelet", Wseas Transactions On Signal Processing, Issn: 1790-5052, Issue 5, Vol. 4, Pp. 330-339, May.

[2] Sahoolizadeh H., Abolfazl A.,(2009)," Adaptive Image Compression Using Neural Networks", 5th International Conference: Sciences Of Electronics,Technologies Of Information And Telecommunications, March 22-26.

[3] Yi Xing Z., Zhang Y., Hou Y., Jia L., (2007), " On Generating Fuzzy Systems Based On Pareto Multi-Objective Cooperative Coevolutionary Algorithm", International Journal Of Control, Automation, And Systems, Vol. 5, No. 4, Pp 444-445.

[4] Hasanzadeh R., Moradi H., Sadeghi H.,(2005)," Fuzzy Clustering To The Detection Of Defects From Nondestructive Testing", International Conference: Sciences Of Electronic, Technologies Of Information And Telecommunications, March 27-31, Tunisia.

[5] Nahm, U.Y. and Mooney, R.J. (2008) “Text mining with information extraction.” *Proc AAAI-2009 Spring Symposium on Mining Answers from Texts and Knowledge Bases*. Stanford, CA.

[6] Azizah Jaafar. “Information Retrieval Technology”, Springer, 2014.

[7] Thomas M. “Elements of Information Theory”,2nd Ed. Willey-Interscience, 2006.

[8] David J. C. Mackay, “Information Theory, Inference and Learning Algorithms”, Cambridge University Press, 2003.

[9] John G. Prookis, “Digital Signal Processing” 4th Edition, Prentice Hall, 2006.

[10] Alan V. Oppenhein, “Discrete-Time Signal Processing” 3rd Edition, Prentice Hall, 2009.

[11] Dimitris G. Monolakis, “Applied Digital Signal Processing: Theory and Practice”, Cambridge University Press, 2011.

[12] Nikola Stosic, “Screw Compressors: Mathematical Modelling and Performance Calculation”, Springer, 2010.

[13] Laung-Terng Wang, “System-On-Chip Test architectures: Nanometer Design for testability (Systems on Silicom)”, Morgan Kaufman, 2007.

[14] [14] Daniel Glover, “Compressing Subbanded image data with Lempel-Ziv-based coders”, NASA,1993.

Design of Socket Based on Intelligent Control and Energy Management

Jiang Feng
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Dai Jian
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Wu Fei*
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Zou Yan
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

***Abstract*—Smart home is one of the main applications of internet of things, and it will realize the intellectualization of household. Smart socket is part of the smart home, which can be controlled remotely by power supplied, monitor utilization condition, communication network and other functions. This article mainly introduces the intelligent electrical outlet of each hardware modules; software part mainly analyzes the socket's communication mechanism, and the electricity consumption of collected power statistics through diagrams to feedback through wireless communication. Things achieved in an environment of communication between the user and the smart power outlet timely feedback to the user, so as to achieve energy-saving purposes.**

Keywords—Internet of things; Smart home; Intelligent electrical outlet; Wireless communication; Power statistics

I. Introduction

The fast development of electrical makes our lives become more and more convenient. Social demand for power supply capacity is becoming more and more strictly. On one hand is how to save power; On the other hand is how to meet the need of the society of electricity [1],this area has being one of the most intractable problem throughout the world, under the environment of internet of things intelligent socket realizes to consume capacity and feedback to the client's function timely.

Functions and designing method of every module are introduced in this article in detail. Intelligent electrical outlet is an important part of intelligent household system, it can realize the power equipment monitoring easily, control and transmit data, communication technology uses the WIFI wireless technology; the technology is very suitable for household system communication network [2]. Intelligent socket of the original concept is for the purpose of security and protection, it is on the basis of ordinary socket extending more functions, such as its several protection functions against short circuit, overvoltage and lightning. Now smart socket is on the basis of these features combined with the functions of control circuit remotely, and the statistical power of connection on the socket device, through the network it will transmit the data to the users, this data provides a basis for users make decisions wisely, it will make appliances become more intelligent eventually.

Intelligent socket sets WIFI module internally [3], the client that can perform such operations on an ordinary smart phone through the socket. Through intelligent control can switch implement control other switch remotely, operation results will also return to the remote control. It plays a switch setting's role. Intelligent socket mainly is used for household intelligent functions, it often matches with other electrical appliances. Intelligent sockets [4] can open circuit power supply, so as to control the electrical power supply or not; And intelligent host to communicate with each other to ensure the smart home system work normally. Some of the applications build on real time utility data will serve to user; others will analyze data in aggregate and serve policy and other decision makers. As a device can switch power supply of smart home, intelligent electrical outlet is the main medium of the future smart home life. Intelligent home has emerged with the development of computer science, communications, network and automatic control technology and so on[5]. It is also a great change brought up by advancement of technology.

II. Intelligent Socket Hardware Structure

A. Framework of Design

The internal structure of the socket mainly includes MCU, WIFI communication equipment, memory ,relays, electric measurement module, keys, voltage conversion circuit, protect circuit, the power input port and the power supply output pin, which is shown in Fig.1.The following picture presents the links between the related modules. The core is STM32, which will introduced detailed in the following article.

B. The microprocessor section

Due to using 32-bit embedded microprocessor ARM, STM32F103 is selected as the key controller of the system, this module is the core of the system, and its functions include data acquisition, dynamic data exchange, data operation, data statistics, data storage etc. Part of the MCU diagram as shown in Fig.2.

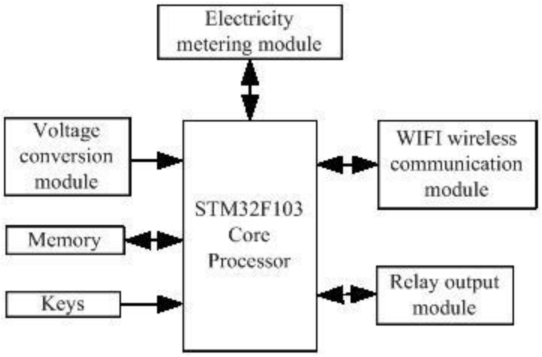

Electricity metering module
Voltage conversion module
STM32F103 Core Processor
WIFI wireless communication module
Memory
Keys
Relay output module

Fig. 1. The Global Structure Of Intelligent Socket

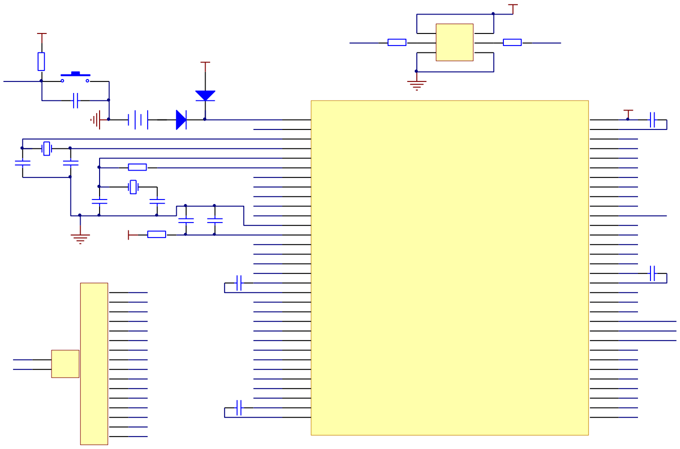


Fig. 2. Part Of The MCU Diagram

To facilitate the observation, a tie-style LED indicator shows the usage status. The circuit diagram is show as Fig.3.

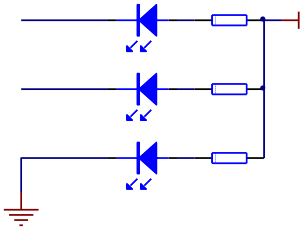


Fig. 3. A Tie-style LED Indicator

The system power supply indicator light is PWR, its light is blue.LED0, and LED1 are connected with PA8 and PD2 respectively. The USB serial port, USB, power supply part of the circuit diagram is shown as the Fig.4.

Fig.4. shows this circuit through the RST and DTR to control BOOT0 and RESET signal, so as to achieve a key download function modules. This section also has switch BUTTON, which is used to control the power of the whole system, if it is disconnected, the whole 3.3 V part of the system will be power outages. And the part 5V power supply is still open. It is shown in Fig.4. F1 is considered as recoverable fuse, for protecting the USB.

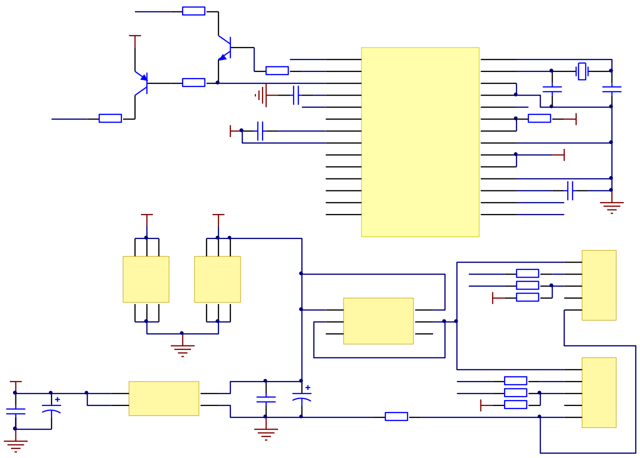


Fig. 4. The System Power Supply Indicator Light

C. WIFI wireless communication section

Based on the design requirements of master controller, we choose the USR-WIFI232—C322 chip [6] to realize hardware design together with other peripheral circuit, it also can use other chips and circuit to realize the WIFI communications functions. The main function is to realize the WIFI communication, to communicate with the MCU on data collection. The circuit diagram is shown in the Fig.5 below.

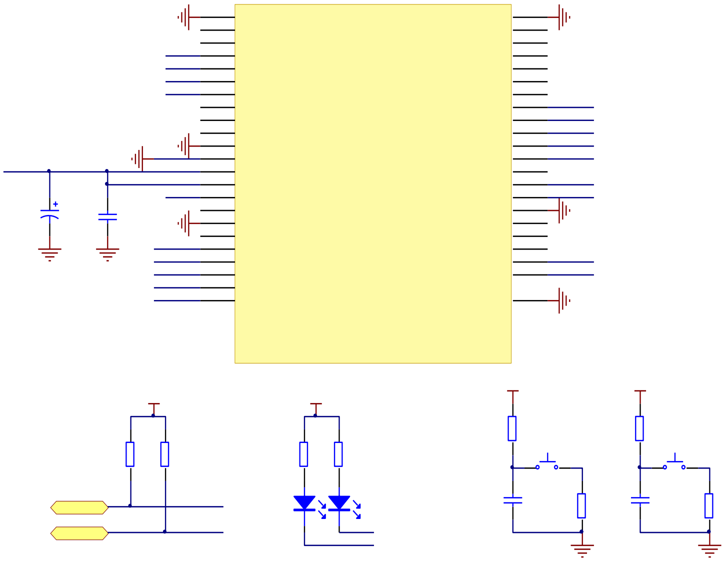


Fig. 5. WIFI Wireless Communication Circuit

D. Memory module

The memory may store WIFI parameters and other data for storage, and sequences of instructions that are performed by the CPU, or any other device are included in the computing system. The storage parameters including WIFI wireless network mode, the connection of the router network name, password, selection of WIFI network IP address, destination IP address, etc.

After the socket makes connection, MCU reads the parameters from memory [7], it will control WIFI to generate WIFI network automatically or connect with a router. MCU reads target parameter information from memory, this system call through WIFI devices to send control code and other data to a specified device.

E. *Relay module*

The Relay connect with the MCU's GPIO port by triode [8], MCU controls relay on or off through the output of the GPIO port, which is shown in Fig.6.

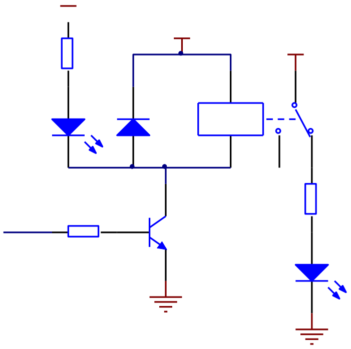


Fig. 6. Relay module

Power supply is available only when the relays are connected. At the same time, on the connect circuit between the relay and triode parallel a led, acting on the indicator. A lighted LED indicates a working relay, however flickers with any activity. It can control socket on or off through STM32 control relays.

F. *Power Measurement System*

It will inevitably contain many sensor modules inside the hardware during intelligent home appliance. The power energy metering module acts as a sensor. It becomes one of the fundamental sensors during many sensors in the home appliance, the following parameters are measured: voltage, current, power factor, power, reactive power and apparent power [9].

Electricity metering module [10] is mainly be composed of a set of EEM301 modules, when the socket is connected, it can realize the inspection of the values about the output current and voltage from the socket, and the results will be sent to the MCU. Electric power management is very necessary, we don't use the simple way of calculation about power is multiplied by the time, which choose the level metering chip, it can satisfy the precision of electricity measurement requirements.

On this part we use EEM301 module for the intelligent power acquisition, it is applied to the smart home, for collecting the electrical power consumption, real-time power, real-time voltage, current, and it can connect WIFI, wireless communication modules by UART, SPI and other wireless communication modules. Users can remote monitor the home appliances by visit the WEB server. So it achieves the remote monitor and intelligent purpose.

In the part of hardware design, voltage and current sampling, the power supply circuit design, circuit communication that is designed. The circuit schematic diagram of the power supply is designed as the Fig.7.Power supply is based on LD0 chips, and takes the form of diode M7 as the protection circuits. By employing the designs, the reliability of the over-voltage protection circuit, the source, the source system, the electronic device is improved.

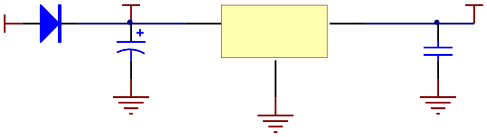


Fig. 7. Circuit Schematic Diagram Of The Power Supply

485 communication circuits use the technique of optical coupling isolation within the input, output and power supply module. From the design of hardware, the power supply is isolated, its requirements is simple, the light coupling communication interface can be directly connected with the TXD , RXD, RD pins in EEM301,in this paper we adopted the following circuit is shown in the Fig.8[11].

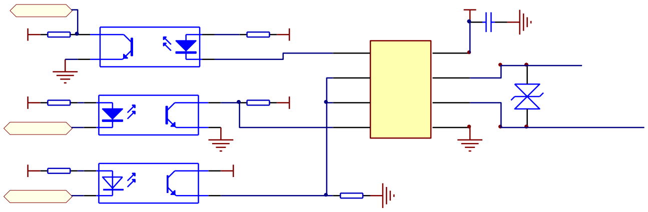


Fig. 8. 485 communication circuit

The current and voltage detection are adopted resistance of sampling and current transformer sampling mode.

G. *The key module*

Key's main function is reset the socket or restore the factory settings. While the push-button is pressed less than 3 seconds, the system will reset when the socket is released then all devices will restart. While the push-button is pressed more than 3 seconds, the outlet will restore the factory default settings. To reset the socket, it will not affect the socket's runtime patterns, connecting network name, password, IP address etc., only to restart; To restore the factory default settings, the socket outlet will work in router mode [12], the socket will generate a WIFI network automatically, through the process of the client it can be connect to the Internet, and it also can set the socket parameters.

H. *Other modules*

Protection circuit is mainly be made of resistance and fuse. When the current is larger than the rated current, the fuse will disconnect automatically in order to ensure safety. If the voltage turns to unwanted fluctuations, and the transient voltage is greater than the specified voltage, through the resistance’s short circuit effect, it will force fuse disconnected to ensure the safety of electricity.

The power input ports are terminal blocks, for connecting wire, zero line and the ground respectively.PE or PEN line are not allowed to connect in series between sockets.

III. The Software Design of the System

The intelligent socket's essence is the socket as a carrier, and it accomplishes a remote smart home system through WIFI wireless [13].That is to say, If we have a super smart socket, no matter where we go, we can control the home air conditioning, lights, television and other household appliances just only with the phone, and we also can realize monitoring of family power and getting its analysis table.

A. The main program flow

The purpose of designing of the system software [8]that is the server mainly be responsible for receiving a serial data from the port, after parsing, according to the inventory build listening socket, the client connection is accepted, maintain the client socket connection, At the same time it will send the received binary data stream to the client. Client application then launches three threads, including connection thread, receiving data thread, data processing thread which they will complete the functions such as data receiving and parsing, displaying and so on.

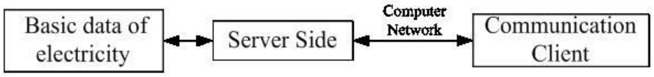


Fig. 9. The physics topology structure between the communication server and its clients

Client applications send a connection request to the server, after setting up the connection the client will send request to the server for reading data, the server processes the request and sends data, the client then starts the corresponding thread for processing. The server establishes TCP/IP connection [14] with MCU firstly, then starts receiving data thread, being responsible for receiving the corresponding data; once the thread has processed the data and displayed then stored it in the database server. To monitor the client connection thread, it is responsible for establishing user socket and listening socket connection.

The system uses the way to connect client applications to the server even in the face of application, platform or network failure. Once a connection is successfully made, the next task often is to read the data thread, read correspondent datum from the buffer, then write the information to a table, avoid the loss of data. View and update data for tables and views: Show as well as update data for tables and views [15].

B. Specific details

For the serial port settings section, first of all, we need to make the serial clock can be set, in APB2ENR registers' configure the serial port 1.Then configure a serial port baud rate in USART_BRR registers. After then, we can go on serial port controlling on the USART_CR1 register. Then in the USART_DR register, data can be sent and received. Of course, a serial port state of nature can be read by USART_SR register.

C. The specific process

After initialing the server which will send out the radio instruction via a serial port at regular intervals, the received instruction socket will reply their own media access control (MAC) addresses to the server.

In general, the server will decode after receiving the response of the frame socket. If the server received the socket for radio command response frame, then obtained the MAC address of the socket [16], and read the output socket's order through a serial port, got the socket’s power information instruction, after receiving the order, then it would send the response frame of power information to the server.

D. The design of the PC client software

The PC client authentication information is divided into three parts, they are user login interface, electricity information query interface, and actual data generated interface. This system establishes log-in boundary wish to realize the identification [17].It is shown in Fig.10.

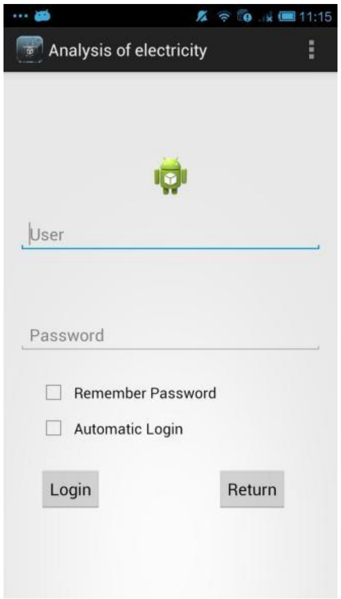
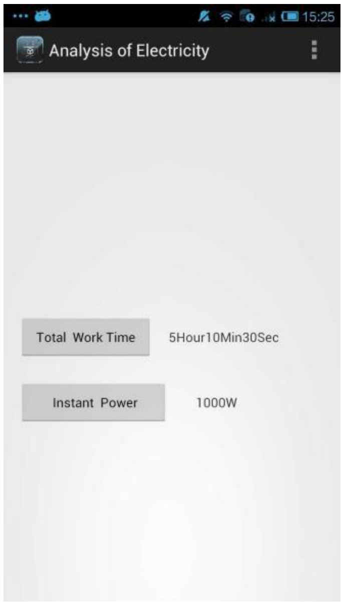


Fig. 10. User login interface

Before being activated, the software read the information of interface elements, and generate graphical user interface (GUI) dynamically according to the information.

The socket's equipment interface can add or delete a socket for users, and it also can control the state of the socket whether open or not. The electricity information that will be displayed in the Fig.11.Show the real-time electrical appliances [18].

The curved line chart draws a full time power. One of the two lines is the actual power, measured by the meter, and the other power is through the power measurement modules which are designed in this paper, As it can be seen clearly from the Fig.12.The two approximately equal, so the feasibility and validity of the molding methods offered [19] in this paper are confirmed in this paper.



15:25

Analysis of Electricity

Total Work Time 5Hour10Min30Sec

Instant Power 1000W

Fig. 11. The real-time electrical information

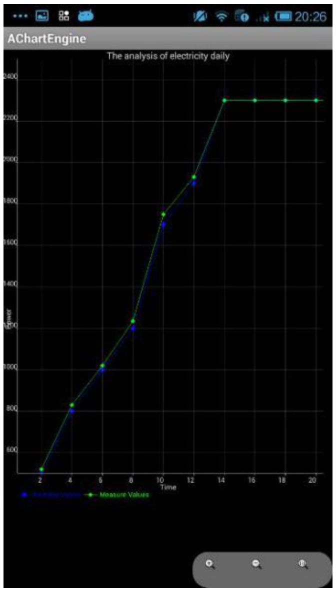


Fig. 12. The real-time electrical appliances

First, Users can accomplish logging the system remotely, and they can view the device’s current work status, ultimately the client will allow the electricity used by domestic appliances to be "dynamically" managed .This design reduces the waste of electricity by users. This will alert users track how much energy each appliance is using and command it to use less. This will alert the user to use a power saving device, so get rid of some bad habits of electricity or energy-consuming devices can be replaced with energy-saving equipment according to the situation.

E. The software design of the intelligent mobile terminal

The major implementation of the smart phone terminal interface is showing the electricity situation, and electrical function. The Intelligent mobile terminally connect to the web via 3G network. The requestor can access the specified URL, an XML aware static reporting program can read that data directly. And it is displayed on the phone’s screen of the result. The users can control the condition of the sockets and devices through the electric state interface. And it also can control the opening and closing of the corresponding socket [20].

IV. Conclusion

The automatic control system realizes the automatic control for intelligent electrical outlet by WIFI wireless communication technology and Internet technology. The system has much advantages in reliability and performance, It will better realize the automation and intelligent in control process.

The next further job mainly realizes the intelligent socket all the way of timing control, it also can add a leakage protection switch to regular socket, eventually it will be able to make it more practical by analyzing and improving them. Later, we can add an LED display in the smart socket further, it can make convenient for users to intuitive interact and experience intuitively. So as to solve a common problem the current socket owned: the user does not know what the current state of the socket presently. However the display screen can solve this problem. Of course, on the other hand intelligent socket also need to keep the traditional control mode, so that the user can according to own needs to choose whether to let it "smart".

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No.61272097, 61405045), the Shanghai Municipal Committee of Science and Technology Project (No.13510501400), Shanghai Municipal Education Commission, A Key Innovation Project (No.12ZZ182).

References

[1] Chen Hai Wang, Zhen Juan Zhang, Ming Huang. The design of the wireless intelligent socket system in the smart home. Journal of TV Technology, vol.37 (2013):80-83.

[2] Yi chao Jin, Lijuan Sun, Ruchuan Wang. The design of the intelligent electrical outlet in Internet environment. The 4th China sensor network academic conference proceedings in CWSN2010.pp.321-326, 2013.

[3] Yichao Jin, Lijuan Sun, Runchuan Wang. The design of the intelligent electrical outlet in Internet environment. Journal of Research and development of Computing machine, vol.47(2010):321-326.

[4] Jin Wang. Research of wireless digital home networks. Xi AN Northwestern university,2010.

[5] Wei Xu, Yuanjian Jiang, Bin Wang. Application of ZigBee technology in the intelligent electrical outlet design. Journal of Communication of Power System,vol.32(2011):78-81.

[6] Fangrong Xu. Design of Wireless Intelligent Home Control System. Journal of Modern architecture electrical, vol,1,2010.

[7] Xianchang Min, Lican Huang. Research of Web service technology based based on the Android platform. Journal of Industrial Control Computer,vol.24(2011):92-94.

[8] Jinhua Peng, Shaolong Shu, Feng Lin, Zhiqiang Huang. A review of several studies on the Home energy management system. Journal of Management of Power demand side,vol.13(2011):35-38.

[9] Hashizume A, Mizuno T, Mineno H. Energy monitoring system using sensor networks in residential houses[C]//26th International Conference on Advanced Information Networking and Applications Workshops (WAINA).Fukuoka:IEEE Press,2012:595-600.

[10] Han J,Choi CS,Lee I.More efficient home energy management system based on Zigbee communication and infrared remote controls[J].IEEE Transactions on Consumer Electronics, 2011,57(1):85-89.

[11] Wenxuan Yao, Zhaosheng Teng, Jingwen Xiong, Yuanning Deng,Tan Xia.Design of Multi-functional intelligent socket. Journal of Technology development,vol.29(2010):28-30.

[12] Huifang Sun, Chundong Mo. Design and implementation of smart home system based on the STM32.Journal of Electronic design engineering,vol.22(2014):82-85.

[13] Shuihong Wang, Wei Xu, Liping Hao.STM32 series ARM Cortex-M3 microcontroller principle and practice.BeiJing:Beijing University Of Aeronautics And Astronautics Press,2008.

[14] Yuesheng Xiang, Duanxia Gao, Yangbo Wu.The design and implementation of intelligent socket based on single chip microcomputer. Journal of Industrial control computer, vol.25(2012):129-131.

[15] Xiaofeng Yao, Chunguang Zhang. The application and JAVA implementation of Socket programming technology in embedded gateway development, Journal of Industrial control computer, pp.21-22,2006.

[16] Fischer C.Feedback on household electricity consumption: a tool for saving energy[J].Energy efficiency,2008,1(1):79-1104.

[17] Lien C H,Bai Y W,Lin M B.Remote-controllable power outlet system for home power management. Consumer Electronics,IEEE Transactions on,2007,53(4):1634-1641.

[18] SAMEK M.Practical statecharts in C/C++:quantum programming for embedded system. Florida:CRC Press,2002.

[19] Chang H H,Lin C L,Lee J K.Load identification in nonintrusive load monitoring using steady - state and turn - on transient energy algorithms[C]//2010 14th International Conference on Computer Supported Cooperative Work in Design(CSCWD).Shanghai,2010:27-32.

[20] Zeifman M,Roth K.Nonintrusive appliance load monitoring:Review and outlook. IEEE Transactions on Consumer Electronics, 2011, 57(1):76-84.

Mobile Arabchat: An Arabic Mobile-Based Conversational Agent

Mohammad Hijjawi
Department of Computer Science
Applied Science University
Amman, Jordan

Hazem Qattous
Department of Software Engineering
Applied Science University
Amman, Jordan

Omar Alsheiksalem
Department of Software Engineering
Applied Science University
Amman, Jordan

***Abstract*—The conversation automation/simulation between a user and machine evolved during the last years. A number of research-based systems known as conversational agents has been developed to address this challenge. A conversational Agent is a program that attempts to simulate conversations between the human and machine. Few of these programs targeted the mobile-based users to handle the conversations between them and a mobile device through an embodied spoken character. Wireless communication has been rapidly extended with the expansion of mobile services. Therefore, this paper discusses the proposing and developing a framework of a mobile-based conversational agent called Mobile ArabChat to handle the Arabic conversations between the Arab users and mobile device. To best of our knowledge, there are no such applications that address this challenge for Arab mobile-based users. An Android based application was developed in this paper, and it has been tested and evaluated in a large real environment. Evaluation results show that the Mobile ArabChat works properly, and there is a need for such a system for Arab users.**

***Keywords*—Conversational Agent; Mobile; ArabChat; Chatterbot and Arabic**

I. Introduction

More than 60 years ago, Alan Turing devised the imitation game (Turing Test) to determine if a computer program could think or imitate the human [1]. Using his game, he tried to prove the machine's capability to act as a human or at least imitating the human through conversations. Turing's game summarised with two separated rooms, in the first room there is a human and in the other room there are a human and a machine. The human in the first room starts text-based natural language conversation with the second party (human or machine) in the second room. The second party (human or machine) will reply to his/her conversations. After certain conversations, the first party (human) should decide who was talking with him either the human or the machine. The game does not check the machine's ability to give the correct answers. Instead, it checks how much the answer close to a human's answer. "Turing's game and people's desire to communicate with computers naturally were the best drivers for the creation of conversational agent" [1].

A Conversational Agent (CA) is an intelligent program that attempts to simulate conversations between the human and machine. CAs can is applicable in many applications such as help desk, information retrieval, education, entertainment, E-commerce and other applications.

From Turing's game time, the researchers compete others to build the most intelligent CA. Their CAs were classified as Embodied CA, Linguistic CA or mixing among them. The embodied CA contains a humanoid character that handle conversations with showing body reactions such as human sounds, facial expressions, and movement of its eyes. Where Linguistic CAs deals with written or/and spoken conversations without to embed the embodied communications.

Different approaches can be used to develop a CA; the most successful common used approach is Pattern Matching (PM). The PM is a technique that use an algorithm to handle user conversations by matching a user's utterance (conversation) against the CA's pre-scripted patterns. These patterns represent the text expression of the expected conversations (sentences). A typical pattern consists of a collection of words, spaces, and wildcards. A wildcard is a symbol used to match a portion of the user's utterance. A number of CAs has been developed using the PM such as in ELIZA [2], InfoChat [3-6], ALICE [7], InCA [8] and ArabChat [9]. To best of our knowledge, there is no Mobile based CA has been developed to handle text-based conversations through mobile devices. Most of the above CAs handle text conversations using a personal computer based application. The following are a brief description for some of those CAs.

ELIZA:

A decade after Turing proposed his test; Joseph Weizenbaum developed a program at MIT to simulate the behaviour of a therapist, called ELIZA [10]. Weizenbaum described it as a program that has the ability to make a natural language conversation with a computer [10]. ELIZA works based on pattern matching, using a few decomposition and re-composition rules. These rules replace some pronouns from the user's utterance with other pronouns and embed them into a response. For instance, it replaces the pronoun "I" from the utterance with pronoun "you" at the suitable position in the response. Consider the following example:

User: My boyfriend made me come here

This user's utterance would match the decomposition rule "my 1 me 2". Where 1 and 2 represent the wildcards that match part of the utterance substrings as follows:

1= boyfriend made

2= come here

Consequently, ELIZA runs the re-composition rule “your 1 you 2”. Where 1 and 2 are the matched substrings in the previous decomposition rule. As a result, the ELIZA’s response is:

ELIZA**:** your boyfriend made you come here

It is possible to notice from the above example that ELIZA tries to ask questions derived from the user’s input to give the impression that it is interested in the user conversation. By using this method, ELIZA tries to keep the conversations going for as long as possible. ELIZA matches the user’s input against a series of stored patterns. If ELIZA finds a match then, it replies with part of the response taken from the user’s input. Otherwise, ELIZA will reply with some fixed responses such as “Very interesting. Please go on” or “Can you elaborate on that”.

ELIZA has many drawbacks. One such drawback, it does not use any logical inference to understand the meaning of the user utterances or even determine the patient’s topic. Instead, it uses simple string searches and manipulation and thus ELIZA has no grammatical knowledge just make pronouns swapping. ELIZA replies by rephrasing many of the patient's statements as questions and posing them to the patient in an attempt to encourage them to elaborate. Therefore, ELIZA’s responses sometimes make a human feel that he/she is speaking to himself/herself. Also, ELIZA cannot keep the continuity of the conversation, and it is not able to store utterances’ information, which might be needed during the same session [11].

ALICE:

Dr. Richard Wallace started developing ALICE in 1995, and continuously improved it over the years [12]. ALICE has won the Loebner Prize three times from the year 2000 to 2004 as the most ‘human-like’ Conversational Agent. ALICE relies on pattern matching technique to handle a user’s conversation, and it uses AIML (Artificial Intelligent Mark-up Language) language to script its knowledge base. The following scripts consider as an example of AIML format [13].

```
<aiml version=”1.0”>

    <topic name=”my topic”>

    <category>

    <pattern> USER INPUT </pattern>

    <that>THAT</that>

    <template>CHATBOT ANSWER</template>

    <category>

    <topic>

    <aiml>
```

AIML, which is a specification of XML (Extensible Mark-up Language), organises the scripted knowledge into AIML files. Each AIML file starts with the <aiml> tag to indicate the AIML used version and also includes a series of AIML units called topics and categories. The ‘topic’ is an optional element (<topic> tag), which has a name, and groups set of categories related to that topic together. Each category (<category> tag) contains a pattern that is matched with the user’s utterance and a template that formulates a reply by that category.

ALICE starts normalising the input before matching it. Normalisation attempts to remove punctuation and convert lowercase letters to uppercase. Then, ALICE starts matching the input word by word depending on the depth first search technique to obtain the best pattern matching. The best matching is the longest pattern match in terms of number of keywords [13]. The depth first search finds the first available matched pattern regardless of the possibility of other pattern matches availability, which means it does not guarantee to select the best match [14]. If there is no match, ALICE follows the ELIZA methodology by asking simple questions on common issues to keep the conversation going.

ArabChat:

ArabChat is a Conversational Agent developed in 2012 to handle conversations for the Arabic language [9]. When ArabChat built, it has been taken into consideration two major factors. The first factor, taking the nature of the human conversation that may start with a specific topic and naturally switch between topics. The second factor is considering the Arabic language morphological nature. Also, sometimes speakers might have to remember some spoken topics or the need to retrieve some parts of spoken utterances and to track the sequencing of subtopics.

The ArabChat uses the PM approach for handling the Arabic textual conversations where the ArabChat’s framework consists of three main components that are the scripting language, engine and brain. The scripting language used to script the applied domain topics to represent them. ArabChat scripting language is a rule-based language, which depends on a rule structure to handle the expected Arabic conversation. ArabChat scripting language can structure any applied domain into a set of contexts (topics), where each context has many rules. A rule (sub-topic) has many patterns and associated responses. While, the ArabChat’s brain is the CA’s knowledge base that is used to store the domain’s scripts where the engine handles the Arabic user’s conversations that target the scripted domain.

The following is an example of a rule in ArabChat:

```
<AlreadyPaid>

a: 0.4

p:15 have paid already.

p:15 *payment is on @ way.

p:15 *have paid already.

p:15 I have @ paid.

r: Ok. That is fine and thank you
```

The rule is to confirm that a payment is already paid. Each rule has a unique name (“AlreadyPaid”), and a decimal value (0.4) called “base activation level” that is used to calculate the rule’s strength. This strength is used by ArabChat to differentiate between competitor-matched rules to select the best one. The rule that has the highest strength will fire. The rule has many patterns that deal with utterances to confirm that the payment was paid. A pattern in ArabChat is a collection of characters, spaces and/or wildcards. Each pattern in the mentioned rule has a base strength (p:15), which it used in the pattern strength matching calculation. Then, the calculated matched pattern strength will be inherited to the rule that this pattern belongs to, to compete with other rules.

After firing a rule, the ArabChat enables a scripter to increase the chance for firing other rules for the next expected utterance by promoting them. Such rules might be related to the fired rule, and they are expected to be targeted by the user after the processed utterance. Promoting a rule means increasing the chance of a specific rule to be fired by increasing its activation level. In contrast, ArabChat can degrade the possibility of other rules being fired (after firing a rule) by decreasing their activation level to the minimum (demoting rules). The ArabChat's scripter can kill rules after firing a rule to prevent them from being fired. Also, ArabChat can manage the navigation between contexts through scripted actions. These scripted actions have the ability to forward the processed utterance to other contexts for further processing or move the agent to another context and wait for the next expected utterance.

InCA:

InCA is an assistant conversational character runs on a handled PDA [8]. InCA uses facial animation and speech input/output to handle user's spoken conversations to provide some services such as appointments and weather reports. In this work, the paper discussed the InCA's architecture with focusing on two limitations of a mobile device platform that are limitation computational power and the input module restriction. The InCA contains three main components; the client that runs on the PDA, the server that manages speech recognition and speech syntheses and the third one is the coordinator who is responsible real-time data retrieval. This conversational agent is not text-based CA, and there are no evaluation experiments on it discussed in the mentioned wok [8].

Most of the above CAs used the Pattern Matching approach for handling the conversations. The Pattern Matching approach showed the impression of some intelligence when it used to handle conversations [15]. The PM approach has many features, including that it is easy to understand, and it is a natural language independent. Also, PM based CAs do not require complex pre-processing stage that might require analysing the natural language sentences that require extra time to process. As a result, the PM is not expensive computationally. Therefore, PM based CAs can handle conversations efficiently for large numbers of users in a real-time environment like the Internet [15]. Moreover, the PM resolves a lot of linguistic challenges such as morphological changes (changes occurring in a word when adding affixes to it). These changes can be resolved using the PM through the pattern's wildcards. Also, the pattern's wildcards can resolve grammatical and spelling errors in a user's utterance. Resolving the spelling and grammatical errors from a user's utterance is an important task that help in continuing the conversations between the user and the CA [16]. From the other side, all considerations that the PM faces are the large number of patterns that the scripter should script to cover the CA's domain. However, the PM approach can minimize a large number of needed pattern to the minimum using variety of wildcards. As a result, this paper chooses the Pattern Matching approach to building the Arabic Mobile CA based on the above supporting features. Also, the Mobile device platform still has limitations in the computational power, displaying and wireless network bandwidth that lead to choosing such a light approach.

II. Need for Mobile Arabic CA

In the last decades, the technology has increasingly evolved in a large spread manner. Researchers tries to automate everything through the evolved technology. This automation has different aspects from different fields such as in education, business, and entertainment and even in social societies. For instance, customers for a company prefer to use an online help desk rather than using the traditional help desk through dealing with the company's employees [17]. Using such online help desk has number of features for customers like availability and cost effective. In addition, it is beneficial for an organisation by reducing the operational costs and documenting all customers' requests for easy analysis and data mining.

One of the most technology trend that has been evolved increasingly in the last decade is mobile smartphones development. Mobile technology has the ability to change the way of people communication and impacting positively in their societies and economies [18]. The world has about 7 billion mobile subscriptions which almost equivalent to the number of people in the planet [19]. This caused to change the way that people dealing with business, learning, socializing and with the government transactions [19]. Arabs are not in isolation from this technology evolution. In the Arabic region, the Mobile communication has expanded rapidly and this can be appeared from the number of mobile subscriptions which has been remarkably increased from 126 million to 350 million in the period of 2006 to the 2011 [20]. According to [19], "For more than a decade, an increasing number of Arab countries have realized that the 'knowledge economy' fuelled by wide adoption and availability of information and communication technologies infrastructure will play an essential role in growth and development". The large number of mobile subscriptions and the 'knowledge economy' led the industry to focus more in smart phones manufacturing and thus in its applications development. 42.07% of Arab users use their smartphones to access the internet instead the personal computer or the laptop and 23.71% use tablets to access the internet [19]. Around 52% of Arab users spending from three to seven hours on the Internet daily from different places like home, work, school, mall and in Internet café [19]. Around 47% of Arab users have from 1 to 25 applications in their smartphones [19]. The smartphone penetration in the Middle East region will grow up to 39% by year 2015 and in Jordan by 50% for the same year [21]. According to [19] and [21], around 40% of Arab users use an android based smartphone.

All of these was the driver to develop the Mobile ArabChat for Arab users in general and for Jordanian people specifically in this paper using the Android. The Mobile ArabChat is an intelligent system that can be used in many applications as mentioned above. In this paper, it has been select the Mobile

ArabChat to work as Information Point advisor for a university's students. However, Arab users still have some challenges in using the Internet. Some of these challenges are shown in Figure 1 [19]. A 48.40% of Arab users found that they suffering from some weaknesses of using the internet in terms of its accessibility and connectivity. In addition, as shown in the figure, 44.68% of users were suffered from the cost and 40.75% are suffering from the lack of the Arabic content on the Internet. Moreover, 30.24% of users faced a limited bandwidth when using the Internet in general. For these reasons, it has been selected to propose and develop the Mobile ArabChat in this paper to handle text-based conversations as the text conversations has small size and limited cost when utilising the bandwidth.

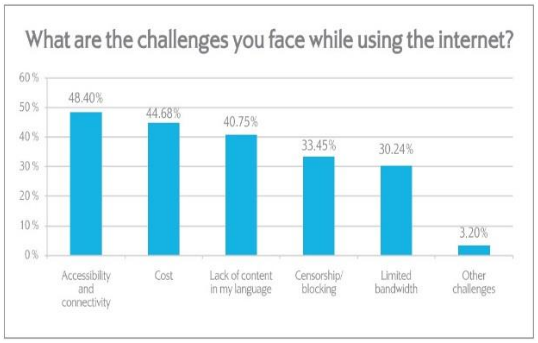


Fig. 1. Arab users challenges when using the Internet[taken from [19]]

III. Mobile ArabChat Framework

The Mobile ArabChat is modelled based upon the ArabChat CA [9]. A new framework has been proposed and developed to issue the Mobile ArabChat to suite the limitations of the mobile device platform. The Mobile ArabChat framework is a rule based CA and contains of number of integrated modules (as depicted in Figure 2) which are scripting engine, scripting language, user interface, temporal memory and knowledge container brain.

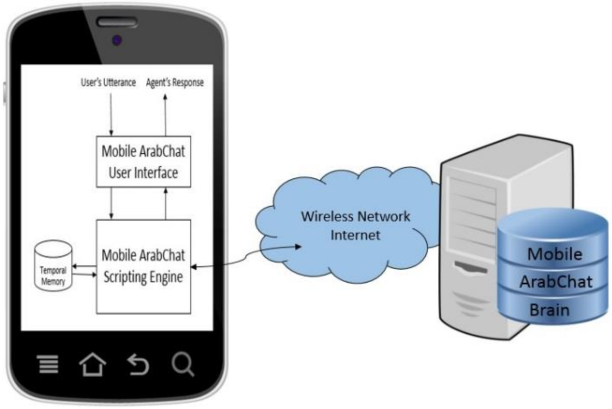

User's Utterance
Agent's Response
Mobile ArabChat User Interface
Mobile ArabChat Scripting Engine
Temporal Memory
Wireless Network Internet
Mobile ArabChat Brain

Fig. 2. the Mobile ArabChat's framework

The Mobile ArabChat scripting engine considered as the core module in this framework. Mainly, it is responsible to handle users' conversations by matching them against the scripted patterns. However, it do number of other tasks such as validating the user's utterance to be sure it is in Arabic and valid before proceeding it to the engine to be processed. In addition, the scripting engine can manage and control the conversations especially when it switch among different topics. Moreover, it can encapsulates the Mobile ArabChat's response with some captured information from the user (when needed). The Mobile ArabChat has a temporal memory to store a captured portion of a user's utterance in order to use it later in the Mobile ArabChat's response which give a good impression at a user. In addition, the Mobile ArabChat can save the processing effort and time and clear an ambiguous when the user targeting the same rule indirect by entering for example "I don't understand you" or something similar by re-firing the previous fired rule with generating another response for the same fired rule. The previous fired rule parameters are stored in this temporal memory and the engine checking it before proceeding with the utterance. Before proceeding with the full framework explanation, it is important to explain the hidden part of this framework which is the Mobile ArabChat scripting language. The following subsections explain main modules of the Mobile ArabChat's framework:

A. *The Mobile ArabChat scripting language*

The scripting language used to script the applied domain topics that covers the Mobile ArabChat service's topic where the engine handles the Arabic user's conversations that target this service's topic. The Mobile ArabChat's brain is the CA's knowledge base that is used to store the domain's scripts.

The Mobile ArabChat uses the PM approach to handle the Arabic textual conversations according to the framework that depicted in figure 2. As mentioned above, the scripting language used to script the applied domain topics in order to represent them. The Mobile ArabChat scripting language is a rule-based language, which depends on a rule structure to handle the expected Arabic conversations. The Mobile ArabChat scripting language structures any applied domain into a set of contexts (topics), where each context has many rules. A rule (sub-topic) has many patterns and associated responses. The Mobile ArabChat scripting language categorises the applied domain's topics through contexts. The domain is the area's topic that the CA will help in such as handling conversations for a company as a help desk employee. Where, a rule is a sub-topic of a context that a user might target in his/her utterance, whereby a pattern is a representation of that utterance which belongs to such a rule. Finally, the responses are the reply to the user's utterance.

In order to handle the conversations, the Mobile ArabChat's scripter should feed/script the applied domain into the Mobile ArabChat's brain categorised as contexts with related rules, patterns and responses. Each context has number of rules to represent the topics inside that utterance. In addition, a context has a "Default rule" to be fired when the utterance targeting the context's topic but no rule in that context has matched that utterance. For instance, if the applied domain was to cover conversations about the tourism places in Jordan, so the contexts might be Archaeological sites, antiquities bazars and therapeutic places such as the Dead Sea. The archaeological sites context might has several rules (sub-topics) such as Petra and Jerash cities. When a user targets tourism in Jordan in general, the engine will reply to the user with a general response about the tourism by firing the rule "default" that belongs to the main context. In contrast, when the user targets the tourism in Jordan in the Petra city, the engine will reply with a response related to the Petra city and according to his/her utterance by firing the “Petra” rule. The Mobile ArabChat offers number of scripting features that manage the scripting process and control the switching between the scripting topics in order to script a coherent domain. The following considers an example of a rule’s structure:

<RegistrationFees>
<رسوم-التسجيل>

a: 0.2
القيمة الإستنادية: 0.2

p: 15 * fees $ registration *
نمط المحادثة: * رسوم $ تسجيل *

p: 15 $ fee * registration #
نمط المحادثة: $ رسوم * تسجيل #

p: 15 $ fee% * registration%
نمط المحادثة: $ %%رسوم * تسجيل *

P: 15 * fees $ registration% *
نمط المحادثة: * رسوم $ تسجيل *

الرد: تبلغ رسوم التسجيل 25 دينار كطلب إلتحاق بالجامعه غير مسترده و 300 دينار أردني كرسوم تسجيل فصلية.

r: The registration fee is 25 JD (Jordanian Dinars), paid once and it is non-refundable and 300 JD registration fees for each semester (term).

The above rule (Arabic scripts and translated English scripts) deals with a sub-topic of an applied domain which concerns of the fees of the university registration for a new student. The rule has number of elements which are; rule name (“Registration-Fees”), base activation level (0.2), number of different patterns (started with p:) and a response (started with r:). Each rule has a unique name and a decimal value called “base activation level” which is used to calculate the rule’s strength. This strength is used by the Mobile ArabChat to differentiate between competitor-matched rules to select the best matched rule. The rule that has highest strength will fire. A rule has many patterns to deal with upcoming utterance by matching it with the scripted patterns. A pattern in the Mobile ArabChat is a collection of characters, spaces and/or wildcards. Pattern’s wildcards that appeared in the above example has different purposes as described below:

Pattern’s wildcards types:

1) The wildcard symbol “$” is used to match or represent one word.

2) The wildcard symbol “*” is used to match or represent many words.

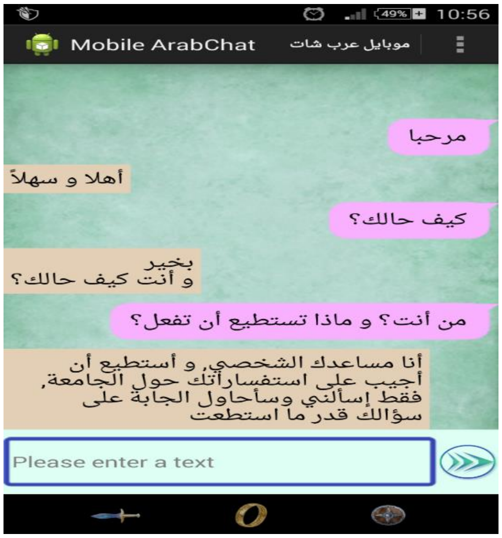
3) The wildcard symbol “%” is used to match or represent one character.

4) The wildcard symbol “#” is used to match or represent one digit.

Each pattern in any rule has a base strength (for example p:15), which it used in the pattern strength matching calculation. Calculation the pattern strength depends on number of factors such as the pattern’s base strength value, number of matched keywords and the length of the user’s utterance. Then, the calculated matched pattern strength will be inherited to the related rule and considered as the new rule’s strength and competing with other rules. After firing a rule, the Mobile ArabChat enables a scripter to increase the chance for other rules to be fired for the next expected utterance by promoting them. Such rules might be related to the fired rule and they are expected to be targeted by the user after the processed utterance. Promoting a rule means increasing the chance of a specific rule to fire by increasing its activation level. In contrast, Mobile ArabChat can degrade the possibility of other rules being fired (after firing a rule) by decreasing their activation level to the minimum (demoting rules). The Mobile ArabChat’s scripter can kill rules after firing a rule in order to prevent them from being fired. In addition, the Mobile ArabChat can manage the navigation between contexts through scripted actions. These scripted actions have the ability to forward the processed utterance to other contexts for further processing or move the agent to another context and wait for the next expected utterance.

B. The Mobile ArabChat User Interface

The Mobile ArabChat user interface manages displaying the conversations among the two conversations parties as appear in Figure 3. The Mobile ArabChat user interface has been developed using the Android technology which it is now holding more than 40% of Arab mobile smartphones [19, 21]. Designing the user interface was simple and user friendly to accomplish its function with the minimum smartphone resources and with a limited Internet bandwidth usage. Each conversation party has different location side in the interface as most of the mobile based chatting application. In addition, each conversation party utterances has different colours as appeared in Figure 3.



مرحبا

أهلا و سهلاً

كيف حالك؟

بخير
و أنت كيف حالك؟

من أنت؟ و ماذا تستطيع أن تفعل؟

أنا مساعدك الشخصي, و أستطيع أن
أجيب على استفساراتك حول الجامعة,
فقط إسألني وسأحاول الجابة على
سؤالك قدر ما استطعت

Fig. 3. The Mobile ArabChat user interface

C. The Mobile ArabChat scripting Engine

The Mobile ArabChat scripting engine has number of integrated functions work together in a novel structure. The Mobile ArabChat scripting engine is the core of this framework and its work according to the PM technique in order to handle users' conversations over the mobile platform.

Once the scripting engine receives an utterance, it starts matching it with patterns of the rules that belong to the current context until all rules of the current context are processed. Initially, the rule that has the highest strength might be fired. However, during the matching process, the engine differentiating between the matched patterns depending on their calculated strengths. The matched pattern that has the highest strength which means having a better match between the pattern and the utterance will inherits its strength to its related rule. Then, the engine will take this calculated pattern strength value and inherited to its rule in order to enable it to compete among other rules in the current context.

In case of no pattern matching the utterance, the scripting engine takes another chance to match the utterance with the patterns of the previously processed context as a precaution step. This precaution step is adopted by the engine to meet the nature of conversations as the consecutive utterance might not have enough explanation in it, by assuming that the previous utterance is already has. If the matching was occurred, the engine will start matching the utterance with the previous context's patterns. Otherwise, a "Default Rule" will fire. A "Default Rule" (DR) is a normal rule but with a higher activation level value (assigned by a scripter) than other rules. A rule that has the highest activation level value means it has a highest strength. A DR's response usually represents a general response for the whole context. Finally, the highest rated rule (regardless if it is a DR or not) will be fired after the conversations' manager is supplied with the needed control information determined by running the fired rule's actions to let the engine moves for which context or to let it remain at the current context and waiting for the next utterance.

D. *The Mobile ArabChat Brain:*

The Mobile ArabChat brain considered the structured store container that the CA's scripter should feed it with the needed structured scripts to prepare the Mobile ArabChat to handle the Arabic conversations. The Mobile ArabChat brain depends on a DBMS (Data Base Management System), to enable a scripter of doing a familiar scripting, searching, querying, and reporting. The scripter should understand the selected domain before start scripting. Then, the scripter start classifying the applied domain into contexts and associated rules. Finally, he/she should script the needed scripts for each rule in terms of the needed variety of patterns and the suitable responses. The Mobile ArabChat offers number of a friendly interfaces to be used by the scripter to script the selected domain. In addition, it offers other facilities such as logs to store the processed conversations and the non-processed conversations in order to track it and analyse it later.

IV. Evaluation

The Mobile ArabChat evaluation methodology is comprised of two main approaches: namely, objective approach and subjective approach. The objective approach will be applied using an light automatic evaluation method called RMUT(Ratio of Matched Utterances to the Total) [9] and manually through analysing the Mobile ArabChat logs. Where, the subjective approach will be performed with recourse to human judgment using the user's questionnaire.

The Mobile ArabChat was deployed on the ASU (Applied Science University) local server and 57 students has been asked to deploy the Mobile ArabChat system into their mobiles handsets. Through the ASU internal wireless network, the students can able to access the Mobile ArabChat. The selected scripted domain is to handle Arabic conversations related to two aspects which are the courses fees of ASU and the total credit hours for each course in ASU. The Mobile ArabChat handled 743 utterances from the 57 users with an average of 13 utterances per user. The following subsections discuss the evaluation from the two approaches side:

A. *The objective approach evaluation*

This evaluation aims to test the Mobile ArabChat performance and the functionality of its main components; mainly the engine. This evaluation will determine whether or not the scripting engine is doing its tasks properly such as recognising patterns' wildcard, matching utterances successfully and the ability to navigate among the scripted contexts. The most challenge in CAs is how to evaluate the CA or to calculate a user satisfaction automatically [14]. However, automating such a task is complex as an utterance has very rich linguistic information especially for a sematic language such as the Arabic.

This evaluation will be done automatically by determining the RMUT (Ratio of Matched Utterances to the Total) of the Mobile ArabChat users. The RMUT is automatically calculated per user by the Mobile ArabChat once a user session is closed and it is automatically calculates the ratio of number of matched utterances to the total utterances for each user (a user's session) based on the following equation [9]:

$$\mathbf{RMUT} = \frac{\textbf{Number of matched utterances}}{\textbf{Total number of utterances}} \qquad (1)$$

The evaluation results show that the average RMUT for the 57 users of the Mobile ArabChat is 83.2%. This means that around 83% of users' conversations has been matched. Initially, this result might be good but for most accurate result a manual analysing for the Mobile ArabChat logs must be done. The meaning of a matched utterance does not always equal to a successful response to a conversation. When the Mobile ArabChat matching an utterance and fired a related rule, the conversation can be considered as a success conversation. But when it matched an utterance and fired unrelated rule, the conversation consider unsuccessful as the reply to the user was incorrect. This conflict might be occurred when two rules shared the same keywords or shared part of the scripted topic. According to [9], the RMUT is not expected to give a full picture about the user satisfaction, but it used to test whether or not the scripting engine is performing its tasks properly in terms of its ability of matching utterances successfully and gives overview of the quality of scripts. On the other hand, it gives a general overview of scripting engine's performance [9].

After the manual analysing job for the logs has been finished, it has been revealed that the actual percentage of successful conversations is 78.64%. This percentage is more accurate and this is caused due to firing unrelated rules or speaking with the Mobile ArabChat outside the scripted domain and that's appeared after the manual testing. However, this reducing in the percentage does not mean any fault in the engine's work but it is clear for the need of an extra effort in the scripting process itself and a user's commitment to speak inside the selected topic. Given this, it can be conclude that the Mobile ArabChat achieved a reasonable performance by its ability of handling/automating the Arabic conversations successfully.

B. The subjective approach evaluation

The subjective approach evaluation will be performed with recourse to human judgment using the user's questionnaire. The evaluation will be conducted by asking Mobile ArabChat users to give their opinion about various aspects of using it. In addition, the evaluation aims to ask the users about their satisfaction of using such a mobile based CA instead of using a desktop based CA. The subjective evaluation aims to enable users to evaluate the Mobile ArabChat user interface, usability, naturalness, the applied domain coverage, speed, availability of similar mobile based Arabic agent, and user general satisfaction. The questionnaire has 15 questions designed to meet the above mentioned evaluation aims. For each aim, a number of questions have been assigned to determine the user opinions concerning them. For each question in the questionnaire, a user has 3 options from which to select his/her degree of approval or disapproval for the asking issue. These options are “موافق” (“Agree”), “محايد” (“Neutral”), “غير موافق” (“Disagree”). The following are the questionnaire questions:

- “واجهة النظام كانت مناسبة جدا” “The user interface was suitable”.
- “كان النظام قادر على إجابتك على جميع إستفساراتك” “The agent was able to answer all your utterances”.
- “أجوبة النظام كانت واضحة ومفهومة.” “The agent responses were clear and understandable”.
- “لم تواجهك أية مشاكل فنية عند إستخدامك النظام” “You experienced no technical problems whilst using the agent”.
- “الوقت المستغرق من النظام للرد على استفساراتك كان مناسبا” “The elapsed time taken by the agent was reasonable”.
- “ تفاعل النظام معك كان واقعي وحقيقي شبيه بتفاعل الانسان من حيث الأجوبة وردود الفعل” “The interaction with the agent was realistic and believable”.
- “ صعوبة التخاطب مع الجامعة عبر الهاتف والبريد الالكتروني وصعوبة الوصول لمعلوماتك المطلوبة عبر موقع الجامعة الالكتروني جعلك تلجأ لإستخدام هذا النظام” “The difficulty of contacting the university by phone or email, and accessing your needed information on the university website were the reasons to use the Mobile ArabChat”.
- “لقد ساهم النظام في توفير جهدك و وقتك .” “The agent saves you time and effort”.
- “ لايوجد خدمة مثيلة باللغة العربية لأي جامعة ,كلية, أو لشركة و أيضا لايوجد نظام اسئلة وأجوبة باللغة العربية عبر الموبايل” “There is no Arabic university, college or company offering the same services over the Mobile platform, even there is no question answering system in Arabic for Mobile users”.
- “كان النظام يشجعك بالاستمرارعلى الحديث” “The agent encourages you to carry on with the conversation”.
- “تقييمك الإجمالي للنظام بأنه ممتازا” “Your overall rating for this service is excellent”.
- “سوف تنصح أصدقائك باستخدام هذا النظام” “You will recommend your friends to use the Mobile ArabChat system”.
- “ أنت تفضل استخدام هذا النظام بدلا عن التحدث مع الشخص المسؤول في الجامعة” “You prefer to use Mobile ArabChat rather than speak with a human advisor”.
- “سوف تعيد إستخدام النظام في المستقبل” “You will re-use this service in the future”.
- “ هل تفضل استخدام هذا النظام عبر هاتفك الذكي عن استخدامه مثيل له عبر جهاز الحاسوب؟” “Do you prefer to use the Mobile ArabChat than using such a system on a personal computer”

1) ArabChat PH2 subjective evaluation results

All users have been answered all questions and submitted their questionnaire and the results as shown in Table 1.

TABLE I. THE MOBILE ARABCHAT QUESTIONNAIRE RESULTS

Question number	“Agree” distribution (Percent)	“Neutral” distribution (Percent)	“Disagree” distribution (Percent)
1	53 (92.9%)	2 (3.5%)	2 (3.5%)
2	47 (82.4%)	7 (12.2%)	3 (5.2%)
3	41 (71.9%)	8 (14%)	8 (14%)
4	56 (98.24 %)	1 (1.75%)	0 (0%)
5	54 (94.7%)	2 (3.5%)	1 (1.75%)
6	15 (26.3%)	7 (12.2%)	35 (61.4%)
7	57 (100%)	0 (0%)	0 (0%)
8	51 (89.47%)	5 (8.77%)	1 (1.75%)
9	57 (100%)	0 (0%)	0 (0%)
10	33 (57.9%)	20 (35.1%)	4 (7%)
11	42 (73.68%)	12 (21%)	3 (5.2%)
12	41 (71.9%)	7 (12.2%)	9 (15.8%)
13	52 (91.2%)	5 (8.77%)	0 (0%)
14	51 (89.47%)	4 (7%)	2 (3.5%)
15	55 (96.5%)	2 (3.5%)	0 (0%)

According to the above subjective evaluation results that shown in Table 1, the mobile based user interface was evaluated using the first question and 92.9% of users agreed that the user interface was suitable. Where the Mobile

ArabChat usability was evaluated through three questions in the questionnaire which are 4, 7, and 8. 98.24% of users agreed that they experienced no technical problems while using the Mobile ArabChat. 100% of users agreed that difficulty contacting the university by phone or email, as well as difficulty accessing their needed information on the university website were the reasons that caused them to use the Mobile ArabChat. Finally, 89.47% of users agreed that the agent saved them time and effort.
The naturalness of the Mobile ArabChat has been evaluated through three questions: 3, 6, and 10. 71.9% of users agreed that the ArabChat's responses were clear and understandable. Only 26.3% of users mentioned that ArabChat's interaction was realistic and believable. 57.9% of users disagreed with the notion that ArabChat encouraged them to carry on with their conversation. This inability to encourage further conversations might be due to the response scripting, which fails to encourage users to continue conversations after firing certain rules.

The applied scripted domain for Mobile ArabChat is simple and it is used only to test the CA. However, 82.4% of users (question number 2) agreed that Mobile ArabChat was able to provide all of their requested information regarding the two covered scripted topics as mentioned above. Regarding the Mobile ArabChat interaction speed the interaction speed of ArabChat has been evaluated through item number 5. 94.7% of users agreed that the elapsed time taken by the Mobile ArabChat to handle their utterances was reasonable. Where the availability of similar Mobile CAs was evaluated through item number 9. All users agreed that there is no Arabic university, college or company offering the same service. This high percentage carries a meaning behind it which the Mobile ArabChat might be considered the first Mobile CA responsible for handling user utterances in the Arabic language. The general satisfaction of the Mobile ArabChat users was evaluated through item numbers 11, 12, 13, and 14. 77.2% of users agreed that their overall rating for ArabChat was excellent, while 71.9% agreed to recommend Mobile ArabChat to their friends. 91.2% of users prefer to use ArabChat rather than speak to a human advisor and 89.47% of users confirmed they would use ArabChat for future needs. Finally, Most of users (96.5% of them) prefer to use the Mobile ArabChat instead of using such a system via their personal computers. This means that the good Mobile penetration ratio that affects the Arab countries changed their life of style indeed by depending on their smartphones for all types of communication.

V. Conclusion

This paper has discussed the mobile based Arabic Conversational Agent called Mobile ArabChat. The Mobile ArabChat framework comprises mainly of a novel scripting engine and a rule-based scripting language structured in a novel way to handle topics of the conversations. Topics of conversation classified into contexts. Each context contains rules that themselves consisted of patterns and associated textual and action-based responses. The Mobile ArabChat handled user's conversations using the pattern's matching technique, by matching the user's utterance against scripted patterns through navigation the utterance into the novel scripts structure. The matched patterns that belong to different rules compete among each other based on their matching strengths. The pattern that has the highest strength will inherit its strength value to its rule and thus it will be fired. The Mobile ArabChat differentiates between matches to select the best match (the rule that has the highest strength) that represent the conversation's topic. The Mobile ArabChat has the facility to navigate among topics through the scripted scripts. The applied domain in this paper was simple and for evaluation and testing purposes. However, the Mobile ArabChat shows a good accuracy from the both evaluation approaches sides; objective and subjective. From the objective side, using the RMUT and logs manual analysis showed that it handled well 78.64% of the conversations. This figure might reflect the general user's satisfaction and the Mobile ArabChat's performance. From the other side (the subjective), 73.68% of users who filled the questionnaire agreed that the Mobile ArabChat was excellent. Also, 96.5% of users found that using the Mobile ArabChat on their smartphones better than using the same system through their personal computers. This due to the flexibility of the mobile platform.

Acknowledgment

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

References

[1] TuringFIT Facebook group, A., Computing machinery and intelligence. Mind, 1950: p. pp 433-60.

[2] Weizenbaum, J., ELIZA: A computer program for the study of natural language communication between man and machine. Communications of the ACM, 1966. 10: p. 36-45.

[3] ConvAgent. Business Rule Automation Knowledge Management Conversational Agent. 2005 [cited; Available from: www.convagent.com/what-we-do.html.

[4] ConvAgent. ConvAgent Foundation- ADAM Conversational Agent. 2015 [cited; Available from: www.ConvAgent.com.

[5] Sammut, C., Managing Context in a Conversational Agent. Electronic Transactions on Artificial Intelligence, 2001.

[6] Sammut, C. and D. Michie, InfochatTM Scripter's Manual, Convagent Ltd. . 2001: Manchester.

[7] Wallace, R. ALICE: Artificial Intelligence Foundation Inc. . 2008 [cited; Available from: http://www.alicebot.org.

[8] Kadous, M. and C. Sammut, InCA: A Mobile Conversational Agent. Trends in Artificial Intelligence, 2004. Volume 3157: p. pp 644-653.

[9] Hijjawi, M., et al. ArabChat: An Arabic Conversational Agent. in Computer Science and Information Technology (CSIT), 2014 6th International Conference on. 2014. Amman, Jordan.

[10] Weizenbaum, J., ELIZA: A computer program for the study of natural language communication between man and machine. Communications of the ACM., 1966. Vol 10.: p. PP 36-45.

[11] Block, N., The mind as the software of the brain, in An Invitation to Cognitive Science, D.N. Osherson, et al., Editors. 1995, MIT Press.

[12] Wallace, R.S. ALICE: Artificial Intelligence Foundation Inc. 2008 [cited; Available from: http://www.alicebot.org.

[13] Abu Shawar, B. and E. Atwell, A Comparison Between Alice and Elizabeth Chatbot Systems. 2002. p. 21.

[14] Goh, O., A framework and evaluation of Conversational Agents, in Information Technology. 2008, Murdoch University.

[15] Timothy, B. and G. Toni, Health dialog systems for patients and consumers. J. of Biomedical Informatics, 2006. 39(5): p. 556-571.

[16] Maragoudakisa, M., et al., Natural Language in Dialogue Systems, a case study on a medical application. , in Proceedings of Panhellenic

Conference with International Participation in Human–Computer Interaction. 2001: Greece. . p. 197–201.

[17] University of Alberta: Using Online Help Desk Tools to Enhance Client Service and Department Operations in ECAR-EDUCAUSE CENTER FOR APPLIED RESEARCH. 2007.

[18] LLP, D., Arab States Mobile Observatory 2013, p.b.D.L.f.t.G.A. (GSMA), Editor. 2013. p. 68.

[19] The Arab Social Media, G.a.I.P., Mohammed bin Rashid School of Government, The Arab World Online 2014: Trends in Internet and Mobile Usage in the Arab Region. 2014.

[20] Zuehlke, E., In Arab Countries, Mobile Internet and Social Media Are Dominant, but Disparities in Access Remain, in Population Reference Bureau. 2012.

[21] Go-Gulf, Smartphone Usage in the Middle East- Statistics and Trends, in Go-Gulf. 2013.

Personalized Subject Learning Based on Topic Detection and Canonical Correlation Analysis

Zhangzu SHI, Steve K. SHI, Lucy L. SHI
Smart Education Center, National Research Institute of Smart City and Big Data,
North 4th Ring Middle Road, Beijing, China

***Abstract*—To keep pace with the time, learning from printed medium alone is no longer a comprehensive approach. Fresh digital contents can definitely be the complement of printed education medium. Although timely access to fresh contents is becoming increasingly important for education and gaining such access is no longer a problem, the capacity for human teachers to assimilate such huge amounts of contents is limited. Topic Detection (TD) is then a promising research area that addresses speedy access of desired contents based on topic or subject. On the other hand, personalized education is getting more attention because it facilitates the improvement of creativity and subject learning of the students. This paper reveals a patented Personalized Subject Learning (PSL) system that caters for the need of personalized education and efficiently provides subject based contents. An efficient topic detection algorithm for providing subject content is presented. Moreover, since education contents are multimedia based ones with multimodal, PSL introduces Canonical Correlation Analysis (CCA) method to detect multimodal correlations across different types of media. Due to its novelty, PSL has been used as the key engine in a real world application of personalized education system as the smart education module sponsored by a Smart City project.**

Keywords—Topic Detection; Canonical Correlation Analysis; Personalized Education; Subject Learning; Multimodality

I. Introduction

Throughout most of history, only the wealthy have been able to afford an education geared towards the individual learners. For the vast majority, education has remained a mass affair, with standard curricula, pedagogies, and assessments. It has been believed so long as the system insists on teaching all students the same subjects on printed medium in the same way, progress will be incremental. However, now for the first time it is possible to individualize education -- to teach each person what he or she needs and wants to know in ways that are most comfortable and most efficient, which may produce a qualitative spurt in educational effectiveness. How can we improve the performance in education, while cutting costs at the same time? In 1984, it was shown that individualized tutoring had a huge advantage over standard lecture environments: students who received individualized tutoring t performed better than 98 per cent of students from the standard classes. Yet the question is how to make individualized or personalized education affordable. Daphne Koller from Stanford AI Lab [1] argued that technology may provide a path to this goal.

Today timely access to fresh contents is becoming increasingly important in today’s education, and gaining such access is no longer a problem because of the widespread availability of broadband both in homes and businesses. Ironically, high-speed connectivity and the explosion in terms of the volume of digitized textual content available have given rise to a new problem, namely, information overload. Clearly, the capacity for human teachers to assimilate such vast amounts of contents is limited. Topic Detection (TD) has emerged as a promising research area that harnesses the power of modern computing to address this new problem by helping us obtain desired subjects or personalized topics in an automatic way. A topic is defined as a seminal event or content, along with all directly related events and contents . Thus, it is inferred that a topic consists of events and contents, both of which are defined in greater detail [2]. A Topic Detection and Tracking (TDT) is defined as something that happens at a specific time and place, along with all the necessary preconditions and unavoidable consequences. Such an event might be a new movie, an election, or an alien attack. TD enables the automatic discovery of new topics from a news corpus and the subsequent assignment of news documents to the discovered topics [3]. A new topic typically corresponds to a newsworthy incident such as the 2012 US presidential election. Therefore, TD technology is a perfect tool for clustering fresh subject contents.

Moreover, education contents are usually multimedia based. They can be texts, animations, sounds, videos, and so on. Text based TD solution alone is not able to do the final content fusion for personalized contents recommendation. In the process of the cross-media recommendation, the query examples and recommended results need not to be of the same media types. For example, students can receive sound pieces by submitting either an image example or a sound example. This is the so called multi-modal environment. Canonical Correlation Analysis (CCA) is then accommodated to calculate the correlations and measure multi-modality similarities across media types [4].

Despite the fact that existing TD solutions play important roles in their applications [3, 5, 6, 7, 8, 9, 10, 11], they do not explicitly incorporate Language Model and cross-media CCA model into their formulations. Based on previous research [12, 13, 14, 15, 16, 17, 18], a novel personalized subject learning (PSL) system is created based on the above ideas. PSL system is a computer aided education system using TD technologies and CCA methodology. Enlightened by achievements in Information Retrieval (IR) field, Relevance Model (RM) is adopted as the language model for TD. RM is a theoretical extension of statistical language modelling and applicable in both retrieval and TD [19]. By treating education contents as news and stories, both TD and IR methods can be used to retrieve relevant contents and feed them into CCA to analyse cross-media correlations.

The remainder of this paper is organized as follows: In Section 2, key concepts and terms are defined and works directly related to PSL system are reviewed. Section 3 describes a novel approach in terms of TD and CCA. In Section 4, the superiority of the approach of PSL is demonstrated. Finally, in Section 5, conclusions and some future research directions are presented.

II. Diagram and Example of Personalized Subject Learning

Personalized education refers to providing learning experiences tailored to each student's interests and learning styles. It also implies student-directed and self-managed learning. Teachers may individualize instruction in a classroom setting but admit that this is hard to accomplish given the competing need to cover subject matter material. Well programmed computers, whether in the form of personal computers or hand-held devices, are becoming an alternative choice. They will offer many ways to master materials. Students (or their teachers, parents, or coaches) will choose the optimal ways of presenting the materials. Appropriate tools for assessment will be implemented too. Most importantly, computers are infinitely patient and flexible. With a computer aided personalized subject learning system, human beings can spend the precious classroom time on more interactive problem-solving activities, which may help them achieve better understanding and foster creativity. Once the personalized education takes hold, the world will be very different. Many more individuals will receive better education because they will be learning knowledge in ways that suit them best.

A personalized subject learning system consists of 17 components, as shown in Figure 1. Components 1-4 capture various sorts of input from students, including motion, speech, drawing and text input. Component 7 performs efficient topic detection task and contents are fed in from various sources (e.g. Component 11-offline contents such as digital library, Component 12-contents edited by teachers (component 17), Component 13-online contents (learning from books alone is no longer the way to keep pace with the time). Fresh online contents are definitely the complement of printed education medium. Component 9 records learning behaviour of students and stores them into behavioural log. Contents that deserve to be education materials are collected by Component 10. The Component 6 analyses the correlations among collected multimedia contents and recommends the personalized subjects and contents to students and teachers. Component 8 is responsible for relevance feedback based on likes and dislikes of students and teachers as a mean to justify and improve the effectiveness of PSL.

The inputs and outputs of PSL system follow the sequence of Figure 1 To further elaborate theFigure, real examples are shown here in Figure 2 and Figure 3. Topics are detected and clustered as indicated by red arrows as shown in Figure 2 and then CCA decides which topic suits the personal needs of students as shown in Figure 3. In Figure 2, topics with "Old Summer Palace" have been detected and related contents are clustered to feed in the PSL system.

Precious antiques (textual and image contents) from Old Summer Palace have been returned to China as shown in the 1st and the 2nd pictures in Figure 3. The Film "12 Chinese Zodiac" directed by Jackie Chan (video content shown in the 3rd picture in Figure 3) has been co-related as teaching contents by CCA.

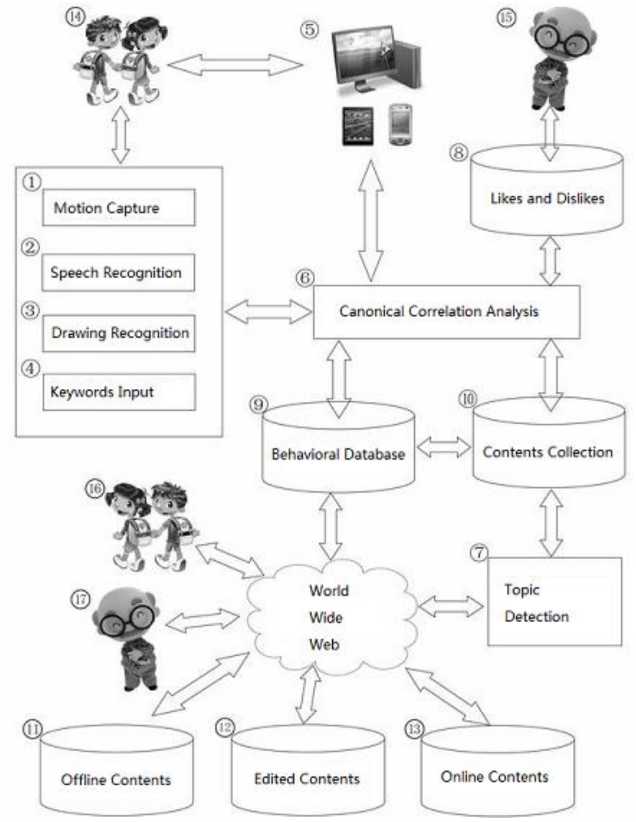
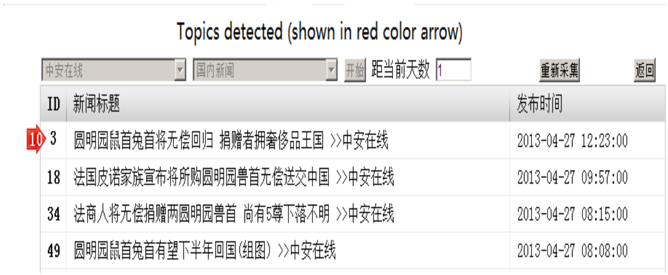


Fig. 1. Framework of Personalized Subject Learning System


Topics detected (shown in red color arrow)

ID	新闻标题	发布时间
3	圆明园鼠首兔首将无偿回归 捐赠者拥奢侈品王国 >>中安在线	2013-04-27 12:23:00
18	法国皮诺家族宣布将所购圆明园兽首无偿送交中国 >>中安在线	2013-04-27 09:57:00
34	法商人将无偿捐赠两圆明园兽首 尚有5尊下落不明 >>中安在线	2013-04-27 08:15:00
49	圆明园鼠首兔首有望下半年回国(组图) >>中安在线	2013-04-27 08:08:00

Fig. 2. Topics detected and clustered that are indicated by the red arrow

The system aims to provide personalized education materials based on subjects or topics. Traditional information retrieval (IR) system is not able to meet such a demand. Hence, this paper proposes a PSL system by accommodating efficient topic detection method and canonical correlation analysis method. The former shoulders the task of fast clustering documents from vast and multiple textual content sources into clustered subjects or topics. The latter is responsible for recommending relevant or co-related contents

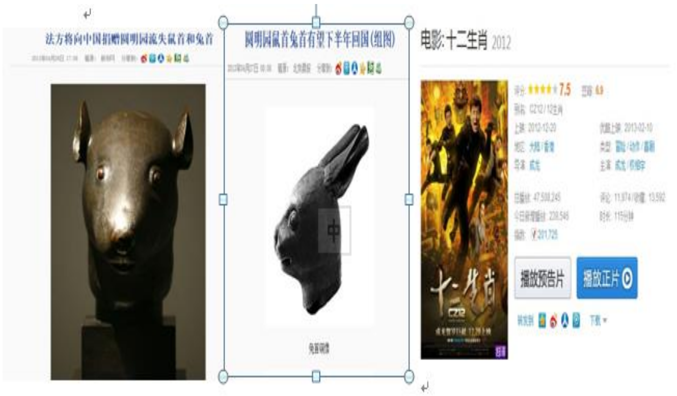

法方将向中国捐赠圆明园流失鼠首和兔首
圆明园鼠首兔首有望下半年回国(组图)
兔首铜像
中
电影:十二生肖 2012
7.5
十二生肖
CZ12
播放预告片
播放正片
下载

Fig. 3. CCA Correlates Antiques Returned to Old Summer Palace and Films by Jackie Chan

by inter-media correlation measure and relevance feedback within the detected topics. Two methods work together to complement to each other for comprehensive, personalized and subject based interactions. Students and teachers then have the easy access to the vast amounts of personalized education contents anytime.

The following section of the paper illustrates the formal representation of two key components, that is, Component 7 for TD and Component 6 for CCA.

III. Formal Representations

In this section, the formal representations of PSL system especially for TD and CCA are described in order. Although there are many language tracking and modelling methods based on machine learning, thus far, the Vector Space Model (VSM) [20] has achieved the best results [21]. VSM has been successfully applied to the well-known SMART text retrieval system [22]. There are a number of formal ways of describing relevance feedback, beginning with the notion of an "optimal query" used in the SMART system. The biggest advantage of VSM is to simplify the text as the vector representation by its features and weights.

A. Document Representation

Contents of the document are expressed by a number of feature items, which generally include the basic linguistic units, such as words or phrases. $Document = D(t_1, t_2, ..., t_n)$, here t_k is a feature item. In a document, each feature item is assigned a weight w_k which denotes the feature item's degree of importance in the document:

$$D = D(t_1, w_1; t_2, w_2; ...; t_n, w_n) \tag{1}$$

Here the weight of t_k is w_k, and $1 \le k \le n$. Given a document $D = D(t_1, w_1; t_2, w_2; ...; t_n, w_n)$, a document can be expressed as a vector of n dimensional vector space. Expression $D = D(w_1, w_2, ..., w_n)$ is called as the Vector Space Model of D. The classic weight calculation method is $TF \times IDF$ in statistical methods. There are many ways to evaluate the significance of a term, ranging from simply identifying its existence to evaluating its distribution level in a document or in a whole corpus. The most common term weighting scheme for processing index terms is $TF \times IDF$, which stands for term frequency — inverse document frequency [21]. $TF \times IDF$ uses the term frequency and inverse document frequency of each feature item to calculate the weight. If tf_{ik} (Term Frequency) represents the number of occurrences of t_k in document D_i, idf_k donates inverse document frequency of t_k, then $TF \times IDF$ is defined as:

$$W_{ik} = tf_{ik} \cdot idf_k \tag{2}$$

Here tf_{ik} is a local statistic value which has different values in different documents. idf_k is a global statistic value reflecting a given term's distribution in all data set. The original definition of IDF is as follows:

$$idf_k = \log\left(\frac{N}{n_k}\right) \tag{3}$$

Here N represents the number of documents in all data sets, n_k represents the number of t_k that appears in data set. It can be seen that, the larger idf_k value is, the less the documents which contain the given term. If all documents contain the same given item, idf_k will be 0. In practice, to avoid such a case, equation (3) is improved by equation (4).

$$idf_k = \log\left(\frac{N}{n_k} + constant\right) \tag{4}$$

Generally, constant value is between 0 and 1, the equation (5) is then induced as:

$$idf_k = \log\left(\frac{N}{n_k} + 0.01\right) \tag{5}$$

If the document length on the impact of weights is taken into account, the feature item weights are normalized into the range of [0, 1]:

$$W_{ik} = \frac{tf_{ik} \times \log\left(\frac{N}{n_k} + 0.01\right)}{\sqrt{\sum_{k=1}^{n}\left[(tf_{ik}) \times \log\left(\frac{N}{n_k} + 0.01\right)\right]^2}} \tag{6}$$

B. TD Representation

The process of topic detection under this model is described here:

1) Topic is defined as $\vec{T} = (f_{T1}, f_{T2}, ..., f_{Tn})$, here $f_{Tj} (1 \le j \le n)$ represents the feature of topic $\vec{T}$;

2) Follow-up story is defined as $\vec{d} = (f_{d1}, f_{d2}, ..., f_{dm})$, and here $f_{di} (1 \le i \le m)$

3) Represents the feature of news story $\vec{d}$;

4) Feature Selection is done by the following two steps:

- Stop words are removed;
- According to descending order of word frequency, the former *i* words are taken as feature items.

5) In TD research field, National Institute of Standards and Technology (NIST) and several universities, including Carnegie Mellon University (CMU), have been established benchmarks and corpus for TDT. In this paper, the similarity between $\vec{T}$ and $\vec{d}$ is defined as follows by adopting the principle reported by Lo and Gauvain of NIST [23]:

$$S(\vec{d},\vec{T})=\frac{1}{L_d}\sum_{w\in d} tf(w,\vec{d})\log\frac{\lambda P(w\,|\,\vec{T})+(1-\lambda)P(w)}{P(w)} \tag{7}$$

Here $S(\vec{d},\vec{T})$ is the similarity of $\vec{T}$ and $\vec{d}$. w is the feature item of $\vec{T}$ and $\vec{d}$. $tf(w,\vec{d})$ is the frequency of w in $\vec{d}$. L_d is the whole number of terms in $\vec{d}$. λ is a smooth factor (0, 1) tuned to make the system achieve minimum cost when tracking TDT3 corpus. TDT3 corpus is created by NIST specially to accommodate Chinese news and stories. The smoothing technique is introduced to prevent data sparsity in unigram modeling.

$P(w\,|\,\vec{T})$ is the probability of w in $\vec{T}$.

$$P(w\,|\,\vec{T})=\frac{C(w,\vec{T})}{Nw(\vec{T})} \tag{8}$$

$C(w,\vec{T})$ is the number of w occurrence in $\vec{T}$, $Nw(\vec{T})$ is the whole number of terms in $\vec{T}$, and $P(w)$ is a priori probability of w which is the statistic value in the background corpus.

$$P(w)=\frac{C(w,background)}{N(background)} \tag{9}$$

Here $C(w,background)$ is the number of w occurrences in background corpus; and $N(background)$ is the whole number of terms in background corpus.

6) According to similarity measurement of NIST, topic detection is then described as the calculation of the similarity between the story and the topic. In other words, if $S(\vec{d},\vec{T})>\theta$, then they are considered as relevant or on-topic, off-topic otherwise.

C. Model Design of TD

Kullback-Leibler divergence is used to compute Relative Entropy (RE) as relevance measure between topic models to compensate the semantic weakness with similar aim of [24].

$$D(M_1\,\|\,M_2)=\sum_w P(w\,|\,M_1)\log\frac{P(w\,|\,M_1)}{P(w\,|\,M_2)} \tag{10}$$

M1 and *M2* are the topic models for topic *T1* and *T2* based on RM**.** The two topic models, *M1 and M2,* both contain the word w . The equation (10) shows whether the two topic models *M1 and M2* have semantic similarity. When value *D* is close to 0, the similarity of two models is high. In order to enhance the robustness of the model, the Clarity probability is introduced for this case when both two models have smaller dissimilarity but they are similar to background corpus [25]. Such a phenomenon is called noise in that it is not a valid topic and therefore should be treated as a noise. Thus, equation (10) becomes the following one:

$$S(M_1\,\|\,M_2)=\sum_w P(w\,|\,M_1)\log\frac{P(w\,|\,M_2)}{P(w\,|\,GE)} \tag{11}$$

Equation (12) is used in the experiment for more convenience of code design and equation (12) is a conversion of equation (11):

$$D(M_1\,\|\,M_2)=\sum_w |\,P(M_1)-P(M_2)\,|+(1-|\,P(M1)-P(GE)\,|) \tag{12}$$

Such a TD model design facilitates code design that then achieves linear performance with the combination of full text retrieval and new algorithm as shown in [16]. Other TD algorithms reported in literature have non-linear performance. The following experiments show lower error rates than those reported in [2].

D. CCA Representation

Content-based multimedia retrieval is a challenging issue, as it aims to provide an effective and efficient tool for searching media objects. Almost all of the existing multimedia retrieval techniques are focused on the retrieval research of single modality, such as image retrieval [26, 27], audio retrieval [28], video retrieval [29] and motion retrieval [30]. However, interactions that enhance students' engagement with Information and Communication Technology (ICT) are multimodal and include gesture, touch, language and so on. Due to the multiple modality of contents, an approach to extend cross-media retrieval to a more generalized multi-modality environment with less manual effort in collecting labeled sample data is needed. In this article, multi-modality representation [31, 32, 33] is adopted as it needs less manual effort in labeling multimedia documents already detected by TD module. In this subsection, the significance appears in inter-media correlation and solution of the problem of heterogeneous topics across different types of medium.

Co-relation of feature space X and feature space Y is defined as follows: $X_{(n\times p)}$ is denoted for n samples and p variables. $Y_{(n\times q)}$ is denoted for n samples and q variables. To obtain the main features, based on their feature weightage, a combination of variables from X and Y is extracted:

$$X_{(n\times p)} \xrightarrow{W_{x(p\times m)}} \mathrm{R}_{(n\times m)};\qquad (13)$$

$$Y_{(n\times q)} \xrightarrow{W_{y(q\times m)}} \mathrm{S}_{(n\times m).}\ (m < p \,\&\, m < q)$$

Here , W_x, W_y are subspace feature vectors. They are supposed to reduce the number of variables and use distribution of *R* and *S* to imitate that of *X* and *Y*. PSL uses relevance coefficient $\rho = r(\mathrm{R},\mathrm{S})$ as in (14) and is optimized by (15).

$$\rho = r(\mathrm{R},\mathrm{S}) = \frac{W_x{}^T C_{xy}\, W_y}{\sqrt{W_x{}^T C_{xx} W_x W_y{}^T C_{yy} W_y}} \qquad （14）$$

is the *covariance* matrix of $X_{(n\times p)}$ and $Y_{(n\times q)}$. Then with Lagrange multiplier method, $C_{xy}C_{yy}{}^{-1}C_{yx}W_x = \lambda^2 C_{xx}W_x$ is computed, which is a generalized Eigenproblem of the form $Ax = \lambda Bx$, and the sequence of W_x's and W_y's can be obtained by solving the generalized eigenvectors. Based on (13), minimum $\mathrm{R}_{(n\times m)}, \mathrm{S}_{(n\times m)}$ is computed to find out the correlation between $X_{(n\times p)}, Y_{(n\times q)}$. For example, let $x_i = (x_{i1},\ldots,x_{ik},\ldots,x_{ip})(x_{ik} \in \mathrm{Re}\,al)$ represents visual feature vector of motion (Component 1 of PSL) and $y_j = (y_{j1},\ldots,y_{jk},\ldots,y_{jq})(y_{jk} \in \mathrm{Re}\,al)$ represents feature vector of speech (Component 2 of PSL). Define x_i by subspace mapping as $x_i' = (x_{i1}',\ldots,x_{ik}',\ldots,x_{im}'),(x_{ik}' = a + b\times i,(a,b \in \mathrm{Re}\,al))$, y_j by subspace mapping as y_j' . Here, subspace is meant for Multi-modality Laplacian Eigen-Maps Semantic Subspace (MLESS).

Due to the existence of large quantity of complex numbers, coordinate values in each dimension of the subspace are converted to their polar form:

$$x_{ik}' = (\beta_{ik,} |x_{ik}'|) \qquad （15）$$

$$\beta_{ik} = arctg(b/a), |x_{ik}'| = \sqrt{a^2 + b^2}$$

The same conversion is done for y_j' . The semantic distance between motion x_i' and speech y_j' is then as follows:

$$CCAdis(x_i', y_j') = sqrt\sum_{k=1}^{m}(|x_{ik}'|^2 + |y_{jk}'|^2 \qquad （16）$$

$$-2\times|x_{ik}'|\times|y_{jk}'|\times Cos\,|\,\beta_{ik} - \beta_{jk}$$

PSL chooses the closest subject coupling with rich media contents and then provides recommends for students and teachers.

Topics are generally clusters of events and contents of specific subjects. To be personalized, clusters need to evolve as students and teachers learn more knowledge and the clusters are also able to optimize the feedback based on his or her experience, opinions, interests and creativity. In this way, personalized education material is finally achieved. In each evolution, the students or teachers have a chance to provide feedback regarding the recommended material and the feedback is treated as a guidance for next TD and CCA tasks.

IV. Experiments

A Java-based personalized education system [43] has been implemented. This system can be easily deployed on any Java virtual machine (JVM) platform.

A. Topic Detection

As a testbed, the system gathered news reports from standard testbed of NIST's TDT3 [2]. Besides, fresh rich media documents from Xinhua News Agency are also added. The experiments tested the viability of our work, in the context of real time fresh online and offline contents of NIST. Detection rate is justified by means of link detection task (LDT) as stated in [16].

B. Relevance Feedback of CCA

By adopting CCA approach co-researched with AI Lab of Zhejiang University, the experimental results of the relevance feedback of CCA [33] fully utilize the contents relevant to the detected topics or subjects, in the context of the user's opinions, creativity, personal knowledge and interests.

C. Practical Deployment

Practical deployment of our algorithm in real world is a patented system in both English and Chinese for personalized education as the smart education module of a Smart City project, as shown in Figure 4. Children's interactions with the computer were frequently referred to, by adult teachers and children, as "playing with the computer" in the same way as they would talk about playing with the bricks or the model animals. The personalized subjects are presented in front of the children as shown in lower portion of Figure 5. This is not surprising inasmuch as the dominant ethos of personalized environments is that children learn through play like a game format [34]: "Children's encounters with books, crayons, and paints were not referred to as play activities, probably because their role in the curriculum was easily identified and practitioners were used to recording children's development in the areas of reading, writing, and drawing. Children's freedom to choose resulted in highly varied patterns of engagement". With same opinions, three categories of teacher involvement, in PSL's computer play, are reactive supervision, guided interaction and a hybrid approach that combines the elements of both.

The application of PSL research investigated learning in personalized settings and an adapted version of the framework and fundamental technology breakthrough have the potential to become research tools and to support changes in practice for professionals in other sectors of education. For example, it is by no means a novel observation that families play a key role in supporting children's learning.

Published during the 1960s, the influential Plowden Report [35] has a section on the importance of parental attitudes and the 'physical amenities' at home. It is recognized that children acquire almost as much general knowledge at the home as in the school, and almost as much information about the world and the way it works during leisure hours as from the formal lessons in the classroom.

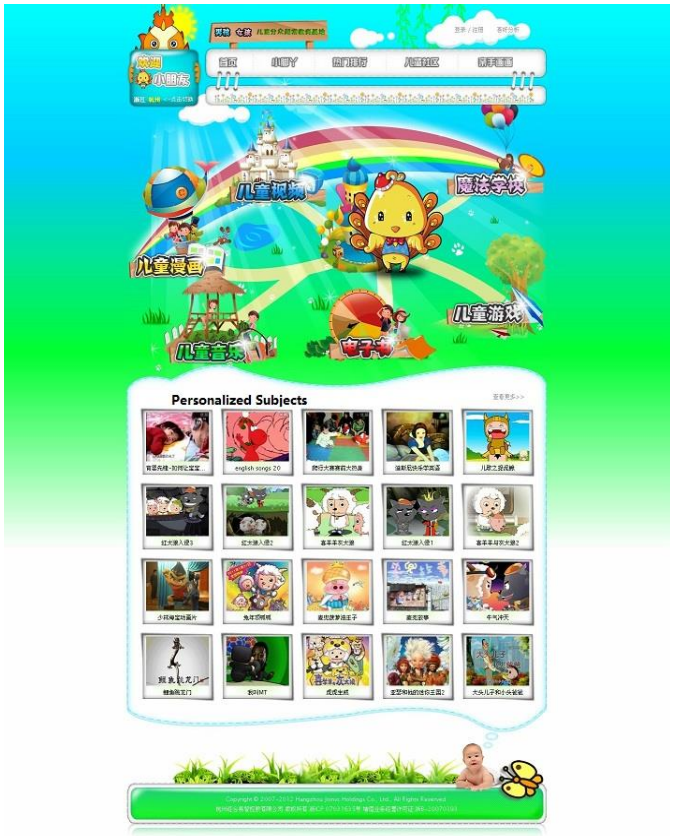


Fig. 4. Personalized multimodal subjects are shown for the students and teachers in the implemented PSL system (smart education module of Smart City project)

Parents can play the role of teachers in PSL since there has been a clear extension in the trends of education from formal settings to the home and more parental engagement [34].

V. Conclusions and Further Works

Due to its efficiency and effectiveness, such a breakthrough meets the practical demands in the fields of Community Question Answering (CQA) [36], social link management [37, 38], learning for personal environment or R&D activities [39, 40, 41], preschool cognitive growth and hence, a distinguished patent has been granted [17].

Think about the guided interaction that helps practitioners to question the purpose of information and communication technology (ICT) and to articulate, reflect on and legitimize the changes in pedagogies. PSL prompts changes in the provision of resources, planning and assessment. Practitioners become more innovative, expand their definition of ICT as well as using existing resources in different ways, and begin to plan for, observe and record student's engagement with ICT in new ways. The breakthrough of PSL in this paper appears not only in the fast TD based clustering but also for CCA based measurable rich media topic recommendation towards subject learning with persistence, engagement and pleasure. Personalized Subject Learning is becoming the trend for people to learn fresh contents. This research shows the capacity and efficiency to automatically deal with vast amounts of information and contents. Hence, the PSL system shows obvious applicability and availability.

In the course of this work, a number of interesting questions have been encountered that we hope to answer in future research. Besides satisfying multimedia contents, the PSL system is able to process multilingual contents in one shot. The research team is currently working on 52 other languages besides English and Chinese. An international PSL system across countries should cater for such a need in the future. It is planned to have in depth collaboration with teams in the States, Europe and Singapore which are keen on PSL and aim to form an international personalized education alliance along with this endeavor.

Acknowledgment

This research is supported by the Chinese National Natural Science Foundation under the grant number of 61073150. The Smart Education Center played a guiding role during the engagement of the Personalized Education Program. The Personalized Education System is sponsored by the National Incubating Center of China and JoinVC Holdings as part of the EASE (Easy Active Systematic Education - 易智童) project, which was initiated under the National Strategy of Smart City [42]. To complement the System, an educational book named "英才是怎样练成的?" is in press, which is supported by JoinVC Holdings, Hangzhou JingWuMen Education Technology and Zhejiang University. The Artificial Intelligence Laboratory of Zhejiang University and Stanford provided us some constructive suggestions in CCA modeling. We would like to express our sincere thanks to all of them.

References

[1] KOLLER, D. , "Technology as a passport to personalized education", page D8 of the *New York Times*, December 6, 2011.

[2] TDT (2004).Topic Detection and Tracking. Annotation manual Version 1.2. http://www.nist.gov/speech/tests/tdt.

[3] BUN, K. K. and ISHIZUKA, M. "Topic extraction from news archive using TF*PDF algorithm," *Proc. of Third Int'l Conf. Web Information Systems Eng. (WISE '02)*, 2002, pp. 73-82.

[4] HARDOON, D. R., SZEDMAK, S. and SHAWETAYLOR, J.. "Canonical correlation analysis; An overview with application to learning methods," *Technical Report CSD-TR-03-02, Computer Science Department, University of London*.

[5] YANG, Y, PIERCE, T. and CARBONELL, J. ."A Study of retrospective and on-line event detection," *Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '98*, 1998, pp. 28-36.

[6] ALLAN, J., LAVRENKO, V. and JIN, H.. "First story detection in TDT is hard," *Proc. of Ninth Int'l Conf. Information and Knowledge Management*, 2000, pp. 374-381.

[7] STOKES, N. and CARTHY, J. " Combining semantic and syntactic document classifiers to improve first story detection," *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '01*, 2001, pp. 424-425

[8] BRANTS, T., CHEN, F. and FARAHAT, A"A system for new event detection," *Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '03*, 2003, pp. 330-337.

[9] KUMARAN, G. and ALLAN, J. "Text classification and named entities for new event detection," *Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM SIGIR '04*, 2004, pp. 297-304.

[10] CHEN, K.-Y., LUESUKPRASERT, L. and CHOU, S. T. “Hot topic extraction based on timeline analysis and multidimensional sentence modelling,” *IEEE Transactions on Knowledge and Data Engineering*, **19**(8), 2007, pp. 1016-1025.

[11] HE, Q., CHANG, K., LIM, E.-P. and BANERJEE A. “Keep it simple with time: A re-examination of probabilistic topic detection models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(10), 2010, pp. 1795 – 1808.

[12] SHI, K., HE, J., LIU, H., ZHANG, N. and SONG, W. “Efficient text classification method based on improved term reduction and term weighting,” *The Journal of China Universities of Posts and Telecommunications*, Vol **18**, 2011, pp. 131-135.

[13] SHI, K., LI, L., HE, J., LIU, H., ZHANG, N. and SONG, W. “A linguistic feature based K-means text clustering method,” *Proc. of IEEE Cloud Computing and Intelligent Systems*, 2011, pp. 108-112.

[14] SHI, K., LI, L., HE, J., ZHANG, N., LIU, H. and SONG, W. ”Improved GA-based document clustering algorithm,” *Proc. of IEEE Broadband and Multimedia Communications*, 2011, pp. 675-679.

[15] SHI, K., LI, L., LIU, H., HE, J., ZHANG, N. and SONG, W. “An improved KNN text classification algorithm based on density,” *Proc. of IEEE Cloud Computing and Intelligent Systems*, 2011, pp. 113-117.

[16] SHI, K. and LI, L. “A Close-to-linear Topic Detection Algorithm using Relative Entropy based Relevance Model and Inverted Indices Retrieval,” *International Journal of Computational Intelligence Systems*, **5**(4), 2012, pp. 735-744.

[17] SHI, K. and SHI, Z. “Subject Shifting based on the Consciousness and Current Focus of Audiences,” Patent, 2012.

[18] SHI, K. and LI, L. “High performance genetic algorithm based text clustering using parts of speech and outlier elimination,” *International Journal of Applied Intelligence*, Vol 38, Issue 4, 2013, pp 511-519..

[19] CROFT, W. B., CRONEN-TOWNSEND, S. and LAVRENKO, V. “Relevance feedback and personalization: A language modelling perspective,” *Proc. of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*, 2001, pp. 49-54.

[20] SALTON, G., WONG, A. and YANG, C.S.”A vector space model for information retrieval,” *Communications of the ACM*, **18**(11), 1975, pp. 613–620.

[21] SALTON, G. and YANG, C.S. ”On the specification of term values in automatic indexing,” *Journal of Documentation*, 29(4), 1973, pp. 351-372

[22] SALTON, G. *Automatic Information Organization and Retrieval*. New York, 1968, NY: McGraw-Hill.

[23] LO, Y. and GAUVAIN, J. “The LIMSI Topic Tracking System for TDT2001,” *Topic Detection and Tracking Workshop*, Gaithersburg, MD, National Institute of Standards and Technology, 2001.

[24] LEE, C., LEE, G. G. and JANG, M. “Dependency structure language model for topic detection and tracking,” *Information Processing and Management***43**(5), 2007, pp. 1249–1259.

[25] LAVRENKO, V., ALLAN, J. and DeGuzman, E. “Relevance models for topic detection and tracking,” *Proc. of the Human Language Technology Conference*, 2002, pp.104–110.

[26] CHANG, E., GOH, K., SYCHAY, G. and WU, G. “CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machine,” *IEEE Transactions on Circuits and Systems for Video Technology*, **13**(1), 2003, pp. 26-38.

[27] HE, X., MA, W. Y. and Zhang, H. J. “Learning an Image Manifold for Retrieval,” *Proc. of the 12th annual ACM international conference on Multimedia*, 2004, pp. 17-23.

[28] GUO,G. and LI, S.Z. “Content-based audio classification and retrieval by support vector machines,” *IEEE Transactions on Neural Networks*, **14**(1),2003, pp.209-215.

[29] FAN, J., ELMAGARMID, A. K., ZHU, X., AREF, W. G. and WU, L. “ClassView: hierarchical video shot classification, indexing, and accessing,” *IEEE Transactions on Multimedia*, **6**(1), 2004, pp. 70-86.

[30] MÜLLER, M., RÖDER, T. and CLAUSEN, M. ”Efficient content-based retrieval of motion capture data,” *ACM Transactions on Graphics*, **24**(3), 2005, pp. 677-685.

[31] WU, F., YANG, Y., Zhuang, Y. and Pan, Y. “Understanding multimedia document semantics for cross-media retrieval,” *Proc. Of the 6th Pacific-Rim conference on Advances in Multimedia Information Processing - Volume Part I, PCM'05*, Berlin, Heidelberg: Springer-Verlag, 2005, pp. 993-1004.

[32] ZHANG, H. and WENG, J. “Measuring multi-modality similarities via subspace learning for cross-media retrieval,” *Proc. of the 7th Pacific Rim conference on Advances in Multimedia Information Processing,PCM'06*, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 979-988.

[33] ZHUANG, Y., WU, F., ZHANG, H. and YANG, Y. “Cross-Media Retrieval: Concepts, Advances and Challenges,” *Proc. of 2006 International Symposium on Artificial Intelligence*. Vol 4261, the series Lecture Notes in Computer Science, 2006, pp 979-988.

[34] Lydia Plowman and Christine Stephen, “Children and computers in pre-school”, *British Journal of Educational Technology*, Vol 36 No 2, 2005 , pp.145 – 157.

[35] CACE - The Central Advisory Council for Education， “Children and their Primary Schools: A Report of the Central Advisory Council for Education (England)”, 1967.

[36] ZHANG, Z. and LI, Q. “Hot topic discovery and trend analysis in community question answering systems,” *Expert Systems with Applications*, **38**(6), 2011, pp. 6848–6855.

[37] GARCÍA-CRESPO, A., COLOMO-PALACIOS, R., GÓMEZ-BERBÍS, J. M. and GARCÍA-SÁNCHEZ, F. “SOLAR: Social link advanced recommendation system,” *Future Generation Computer Systems*, **26**(3), 2010, pp. 374-380.

[38] GARCÍA-CRESPO, A., COLOMO-PALACIOS, R., GÓMEZ-BERBÍS, J. M. and RUIZ-MEZCUA, B. “SEMO: A framework for customer social networks analysis based on semantics,” *Journal of Information Technology*, **25**(2), 2010, pp. 178-188.

[39] COLOMO-PALACIOS, R., GARCÍA-CRESPO, Á., SOTO-ACOSTA, P., RUANO-MAYORAL, M. and JIMÉNEZ-LÓPEZ, D. “A case analysis of semantic technologies for R&D intermediation information management,” *International Journal of Information Management*.**30**(5), 2010, pp.465-469.

[40] GARCÍA-PEÑALVO, F. J., CONDE-GONZÁLEZ, M. Á., ALIER-FORMENT, M. andCASANY-GUERRERO, Mª J. “Opening Learning Management Systems to Personal Learning Environments,” *Journal of. Universal Computer Science,***17**(9), 2011, pp.1222-1240.

[41] GARCÍA-PEÑALVO, F. J., ORDÓNEZ DE PABLOS, P., GARCÍA, J. and THERÓN, R. ”Using OWL-VisMod through a decision-making process for reusing OWL ontologies,” *Behaviour & Information Technology*. Vol 33, Issue 5, 2014, pp. 426-442.

[42] SHI, Z. and SHI, K., 易智童- Personalized Education System, http://www.joyscan.com, www.joypond.com, www.joinvc.com, 2015.

[43] SHI, Z. and SHI, K., “英才是怎么炼成的”- How the Elite was Tempered, book in press, December 2015.

Power and Contention Control Scheme: As a Good Candidate for Interference Modeling in Cognitive Radio Network

Ireyuwa E. Igbinosa¹, Olutayo O. Oyerinde², Viranjay M. Srivastava¹, Stanley H. Mneney¹
¹School of Electrical, Electronic and Computer Engineering,
University of KwaZulu-Natal,
Durban 4041, South Africa
²School of Electrical and Information Engineering,
University of the Witwatersrand,
Johannesburg 2050, South Africa

Abstract—**Due to the ever growing need for spectrum, the cognitive radio (CR) has been proposed to improve the radio spectrum utilization. In this scenario, the secondary users (SU) are permitted to share spectrum with the licensed primary users (SU) with a strict condition that they do not cause harmful interference to the cognitive network. In this work, we have proposed an interference model for cognitive radio network that utilizes power or contention control interference management schemes. We derived the probability density function (PDF) with the power control scheme, where the power of transmission of the CR transmitter is guided by the power control law and also with contention control scheme that has a fixed transmission power for all CR transmitter controlled by a contention control protocol. This protocol makes a decision on which CR transmitter can transmit at any point in time. In this work, we have shown that power and contention control schemes are good candidates for interference modeling in cognitive radio system. The impact of the unknown location of the primary receiver on the resulting interference generated by the CR transmitters was investigated and the results shows that the challenges of the hidden primary receivers lead to higher CR-primary interference in respect to higher mean and variance. Finally, the presented results show power control and the contention control scheme are good candidates in reducing the interference generated by the cognitive radio network.**

Keyword—Aggregate interference; cognitive radio; interference management; interference modeling

I. Introduction

In wireless communication, the availability of spectrum has become scarce due to the rapid growth of wireless communication devices. However, the measurement of spectrum shows that spectrum resources are underutilized in terms of time and space [1]. Due to the growing desire for frequency bandwidth, the conventional method for spectrum assignment becomes unsuitable for the course. In other to make use of the underutilized spectrum resource, the cognitive radio (CR) technology was proposed [2-5] because it allows communication between unlicensed users which is also known as secondary user (SU) and the licensed users otherwise known as the primary users (PU) without causing any harmful interference. A CR user can exist side by side with the licensed PU on the basis of non-interference which is also known as the Interweave CR network or the interference – tolerant basis also known as the overlay CR networks [6-8].

On a non-interference basis, the assigned spectrums to the licensed PU are exploited by the SU for transmission without compromising the PU network [9]. In the interference tolerant basis, the CR user splits spectrum allocated to the licensed spectrum with a condition that the CR user would not cause interference which would be harmful to the PU network. If the interference that emanates from CR network to the PU network becomes assertive and vicious, it would become necessary for the CR network to avoid it. Therefore with these features, modelling and analyzing the interference created by the CR networks it becomes necessary to show the degeneration of the primary network and how the CR network can be employed. The interference modeling of CR network in literature is classified into three main groups which include; spatial, frequency-domain and accumulated interference [10-12]. In the spatial distribution of white spaces white is dependent on the conduct of the primary transmitter such as the geographical position and transmit power. The area fraction of white spaces are studied in detailed in [13-15] it was found out in the study that their enormous amount of white spaces in existence. However, it becomes paramount for the CR to apply robust technologies which suits the power and contention control schemes.

The frequency-domain are modeled using a two-dimensional poison point process, details on the frequency domain are emphasized in [16-17]. Another category of the interference modeling is the accumulated interference, the CR user have to give a guarantee that they would not cause harmful interference to the PU. However, if there is no assurance on the interference, it becomes a challenge to persuade the PU to give access to the CR users to utilize its spectrum. However giving assurance on the intensity of the interference the PU can endure is a challenging task in wireless communication [18-20]. Although, a CR user satisfies the constraint Set could still cause excessive interference to the PU when the CR transmits concurrently with another CR which satisfies the CR sensing constraints.

In this work, we studied power and contention control interference management scheme as a good candidate for interference modeling in cognitive radio. We compared results obtained the two management schemes in situations where there is a perfect knowledge of the location of the primary receiver and an unknown location of primary receiver. We proposed a power and contention control interference management mechanism as a successful method in reducing the CR-Primary interference. The rest of this paper is organized as follows; Section II. Related work. Section III. System Model. Section IV. Interference modeling. Section V. Results and discussion and Section VI. Conclusion

II. Related Work

Although there are significant number of researches has been done in interference modeling but only a few work cited in literature with interference modeling employing all interference management schemes. In [21], a host of heterogeneous CR transmitters around the primary receiver was obtained. This host of SU must give guaranteed services to surrounding PUs. The outage probability was used to evaluate the interference caused by the CR network. This was obtained for the underlay and overlay spectrum sharing scenarios [20]. The interference channel was assumed to be a pathloss only channel in [22-23]. The interference channel is assumed to be a pathloss only channel. However, this work was extended by Menon *et. al* [22] this was done by introducing the shadowing and fading. In all the reviewed work in literature, the CR transmitter is assumed to perform its transmission at fixed power without considering power control scheme. However, it is assumed All CR users in the network communicate simultaneously. In this paper we extended the interference modelling by using the power and contention control interference management mechanism. We derived the probability density function (PDF) numerically. In this paper we model the total interference for the CR transmitters in power and contention control interference management scheme. We considered two scenario of the location of the primary receiver. The first scenario we considered when the primary receiver is known and the second scenario when the primary receiver is unknown.

III. System Model

In this work we have considered a system where the location of the primary receiver is unknown to the CR network. From the figure 1, it is shown that the CR transmitters are distributed around the outer circle. Let R represent the radius of the inner circle and L represent the radius of the outer circle. From figure 1, we have assumed that θ is the angle of intersection between d and d_{pr}.The distance between from the primary transmitters to CR transmitter is denoted by d. While d_{pr} represents the distance between primary transmitters pairs. Therefore the distance between CR-transmitter to the primary receiver d_{cp} is denoted as;

$$d_{cp}(d,\theta) = [d^2 + d_{pr} - 2rd_{pr}\cos(\theta)]^{\frac{1}{2}}, d \in [R,L]; \theta[0,2\pi] \qquad (1)$$

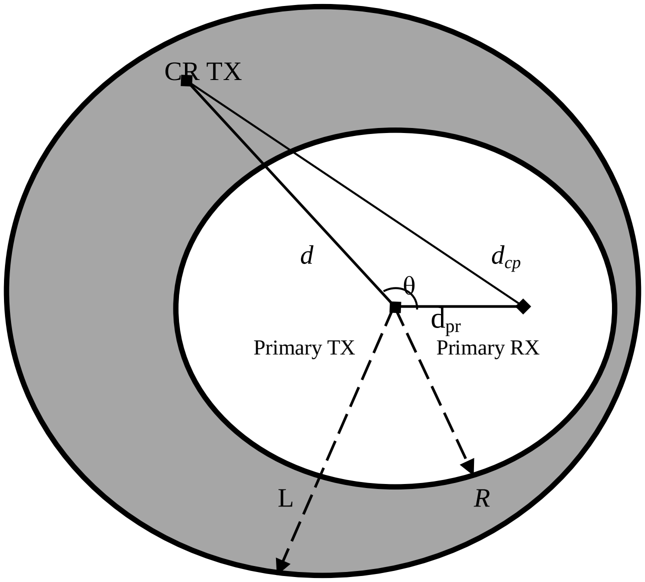


Fig. 1. System model for CR network coexisting with primary network

The CR transmitter is assumed to be distributed following the Poisson point process. Let d be distributed as shown in equation (2); [24].

$$f_d(d) = \begin{cases} 2_x/(L^2 - R^2), & R \le x \le L \\ 0, & otherwise \end{cases} \qquad (2)$$

A. Power Control Scheme

The power of transmission of the CR transmitter is guided by the power control law proposed in [10]. The power control law is represented as;

$$P_{pcr}(r_{ccn}) = \begin{cases} \left(\frac{r_{ccn}}{r_{pcr}}\right)^{\delta} & P_{max,}\ 0 < r_{ccn,} \le r_{pcr} \\ P_{max,} & 0 < r_{ccn} > r_{pcr} \end{cases} \qquad (3)$$

Let r_{ccn} stands for nearest distance from the n^{th} active CR transmitter to its closest transmitter [20]. The power control exponent is denoted with δ while P_{max} represents the maximum transmitting power for CR transmitter. The power control range is represented by r_{pcr} this regulates the minimum r_{ccn} which gives rise to the maximum transmits power of CR transmitter. The PDF of r_{ccn} is represented as;

$$f_{cc}(r_{ccn}) = 2\pi\lambda r_{ccn} e^{-\lambda\pi r_{rccn}^2} \qquad (4)$$

In this work we have assumed that the value of δ and γ which stands for power control exponent and pathloss exponent respectively to be equal, with the boundaries of the power control range the interference is equal to the constant P_{max}or r_{pcr} This is due to the fact that when the power control range is exceeded, the resulting interference is becomes smaller than the constant. Meaning that at any CR transmitter, the interference which emanates from the closest surrounding CR transmitter becomes restricted and independent of the closest neighbor's distance within the neighborhood of the power control range.

B. Contention Control Scheme

In contrast to the previously introduced power control scheme, the contention control scheme has fixed transmission power p at every CR transmitter. However, the CR transmission is controlled by a contention protocol to make a decision on which transmitter can transmit at any point in time [23]. The multiple access protocol in the IEEE 802.11 networks is assumed in this work which is the carrier sense multiple access with collision avoidance (CSMA/CA). All the CR transmitters perform sensing in the medium before transmission, if the CR transmitter identifies a transmission from any CR transmitters around its contention region, it delays its transmission else it starts its transmission. Due to the contention control all CR transmitters are spaced from one and other with a contention distance d_{min} across two CR transmitters. The distribution of all the transmitting CR transmitters can be modelled in a Matern Hard-core (MH) point process [30]. However, approximation for MH point process usually disregards the dependence amongst the CR transmitters and treats an MH point's process as a result of independent thinning process. Therefore all CR transmitters follow the original Poisson point process, with intensity λ however the n^{th} CR transmitter has a probability q_{mh} to transmit at a power level p. The characteristics function of accumulated interference of contention control is given as;

$$\varphi_I(\omega) = exp\left(\lambda\pi q_{mh} \int_H f_h(h) T(\omega ph) dh\right) \tag{5}$$

The PDF of the interference is derived from (5) and (12), however, this is further reduced in similar way like (12) with the following equation;

$$k = q_{mh} \int_H f_h(h) \sqrt{ph}\, dh \tag{6}$$

The detailed derivation of (5) is given in Appendix A.

IV. INTERFERENCE MODELING

In this section, we have modeled the aggregate interference from all CR transmitters by implementing the two interference management schemes which were introduced in section II. We employed the method used in [27] which was later developed by *Hong et. al* in [27], to derive the PDF. This was modeled considering a scenario where the Primary receiver location is unknown to the CR network.

A. Power Control Scheme in when primary receiver location is unknown

From the system model, we have adopted the characteristics based function used in [25-26] and obtained the characteristics function $\varphi_I(\omega)$ of the total interference I at the primary receiver from all cooperating CR transmitters.

$$\varphi_I(\omega) = exp\left(\lambda\pi \int_H f_h(h) \int_p f_p(p) \mathrm{T}(\omega ph) dp\, dh\right) \tag{7}$$

Where $f_p(.)$The PDF of the transmit power of $P_{pcr}(r_{ccn})$ of a CR transmitter shown in (3), the we can rewrite the equation as follows;

$$T(\omega ph) = R^2\left(1 - e^{i\omega g(R) ph}\right) + i\omega ph \int_0^{g(R)} [g^{-1}(t)]^2 e^{i\omega tph} dt \tag{8}$$

From equation (8), $g^{-1}(.)$ symbolized the inverse function of the $g(.)$in the pathloss function in (9)

$$g(r_n) = r_n^{-\gamma} \tag{9}$$

In (5), p is a function of r_{ccn} as expressed in (3). Therefore, the assumption of $T(\omega ph)$ in respect to p is equal to the prediction of $T\big(\omega P_{pcr}(r_{ccn})h\big)$ over r_{ccn}. However, using the PDF which was given in (4) then we can rewrite equation (5) as;

$$\varphi_I(\omega) = \exp\left(\lambda\pi \int_H f_h(h) \int_{r_{ccn}} f_{ccn}(r) T(\omega P_{pcr}(r_{ccn})h)\right) drdh \tag{10}$$

However, (10) can still be rewritten as shown in (11), a detailed derivation of (10) is given in appendix B.

$$\varphi_I(\omega) = exp\Bigg\{\lambda\pi \int_H f_h(h) \int_0^{r_{pcr}} f_{ccn}(r) \Bigg[R^2\left(1 - e^{\frac{i\omega r^{\delta} P_{maxg(R)h}}{r_{pcr}\delta}}\right) + \frac{i\omega r^{\delta} P_{max} h}{r_{pcr}\delta} \int_0^{g(R)} t^{\frac{-2}{\gamma}} e^{\frac{1\omega t r^{\delta} P_{maxg(R)h}}{r_{pcr}\delta}} dt\Bigg] drdh + \lambda\pi \int_H f_h(h) \int_{r_{pcr}}^{\infty} f_{ccn}(r) \Big[R^2\big(1 - e^{i\omega g(R) P_{max} h}\big) + i\omega P_{max} h \int_0^{g(R)} t^{-\frac{2}{\gamma}} e^{i\omega t P_{max} h} dt\Big] drdh\Bigg\} \tag{11}$$

The PDF of the interference is obtained by calculating the inverse Fourier transform on (12);

$$f_I(y) = \frac{\pi}{2\pi} \int_{-\infty}^{+\infty} \varphi_I(\omega) e^{-2\pi i\omega y}\, d\omega \tag{12}$$

The equations (11) and (12), acts as generic statements for the characteristics function and PDF respectively of the interference when implementing the power control scheme. We choose to make use of the value of the pathloss exponent and the power control exponent γ and δ respectively to be equal capping it at value 4. This is because the power control is designed that the interference caused by the n^{th} active transmitter within a power control range is equal to a constant. When the constant is above the power control range the interference becomes smaller. The radius of the interference region R was set as 0. Then the PDF $f_I(y)$ can further be reduced following the steps used by Sousa et. al. in [26] and obtained the following equation.

$$f_I(y) = \frac{\pi}{2} K\lambda y^{-3/2} exp\left(\frac{-\pi^3\lambda^2 k^2}{4y}\right) \tag{13}$$

Let K be the following;

$$K = \sqrt{P_{max}} \int_H f_h(h)\sqrt{h}\, dh \left[\int_0^{r_{pcr}} 2\pi\lambda_e^{-\lambda\pi r^2} \left(\frac{r}{r_{pcr}}\right)^{\frac{\delta}{2}} dr + e^{-\lambda\pi {r_{pcr}}^2}\right] \tag{14}$$

Further derivation of (14) is given in Appendix C

Considering the system model which is depicted in figure 1, and the proposed power control scheme which has been proposed in the earlier section of this work, we can then derive the total interference as:-

$$\varphi_I(\omega) = \lim_{i\to\infty} exp\Bigg\{\lambda \int_H f_h(h) \int_0^{r_{pcr}} f_{ccn}(x) \int_0^{2\pi} \int_R^L e^{i\omega\left(\frac{r}{r_{pcr}}\right)^{\delta} P_{max}(x) g(d_{cp}(d,\theta))^h} d - d\, dr\, d\theta\, dx\, dh + \lambda \int_H f_h(h) \int_{r_{pcr}}^{\infty} f_{ccn}(x) \int_0^{2\pi} \int_R^L e^{i\omega P_{max}(x) g(d_{cp}(d,\theta))^h} d - d\, dr\, d\theta\, dx\, dh\Bigg\}. \tag{15}$$

Detailed derivations of (15) see Appendix D

Interference modelling is known to be computationally complex. It is a known fact that the closed form expression cannot be used for characteristics based functions or Inverse Fourier transforms. However, it is advantageous to model the interference with less complexity. To solve the issue of the complexity of interference modeling, it would be desirable to make an estimation of the PDF of the interference. Therefore in this work we would fit the total interference under the power control and contention control scheme to be log-normal distribution. In [20], it was shown that the sum of interference from evenly distributed interferers in a circular area is asymptotically log-normal. This implies that the total interference why implementing both the power and contention control scheme can be estimated as a log-normal distribution. Also the summation of the randomly weighted log-normal distribution variable can be modeled as a normal log-normal distribution which gives assurance that the total interference is log-normally distributed even when the shadow fading effect is considered in (2) [22]. We have used the cumulant-matching method to approximate the mean and variance of the log-normal distribution function. The randomly distributed variable x was approximated using the first two order cumulant in [10]. The PDF of the log-normal variable x is shown in equation (16)

$$P_n(x) = \frac{1}{\sqrt{2\pi\sigma x}} exp\left\{\left(\frac{-\ln(x)-\mu)^2}{2\sigma^2}\right)\right\} \tag{16}$$

The mean μ and variance σ^2 can be calculated by using its first two cumulant K_1 and K_2 as expressed below [32].

$$\mu = \ln\frac{k_1}{\sqrt{\frac{k_2}{k_1^2}+1}} \tag{17}$$

$$\sigma^2 = \ln\left(\frac{k_2}{k_1^2}+1\right) \tag{18}$$

Taking the interference distribution into consideration, the n^{th}cumulant k_n of the total interference I, can then be derived from its characteristics function $\varphi_I(\omega)$ using the following equation.

$$k_n = \frac{1}{i^n}\left[\frac{\partial^n \ln\varphi_I(\omega)}{\partial\omega^n}\right]_{\omega=0} \tag{19}$$

Using equation (19) and the characteristics function in (11) the cumulant for the total interference when the power control scheme is implemented can then be obtained as:

$$k_n = \frac{2\lambda\pi P_{max}^n e^{n\mu+\frac{n^2\sigma^2}{2}}}{(n\Upsilon-2)R^{n\Upsilon-2}}\left[\frac{n\delta(n\delta-2)\ldots 2}{{r_{pcr}}^{n\delta}(2\pi\lambda)^{\frac{n\delta}{2}}}\left(1-e^{-\lambda\pi_{rpcr^2}}\right)-\sum_{i=1}^{\frac{n\delta}{2}-1}\frac{n\delta(n\delta-2)\ldots.(n\delta-2i+2)}{(2\pi\lambda r_{pcr^2})^i}{r_{pcr}}^{n\delta-2i}e^{-\lambda\pi r_{pcr^2}}\right] \tag{20}$$

This approximation method is applicable to both the pathloss only and shadow fading channel.

When the log-normal approximation is applied to equation (15), we can then derive the k^{th} cumulant of the interference as follows;

$$\begin{aligned} k_n = \lim_{i\to\infty}\lambda\Bigg\{&\int_H f_h(h)\int_0^{r_{pcr}} f_{ccn}(x)\int_0^{2\pi}\int_R^L \frac{d^\delta P_{max}(x)g(d_{cp}(d,\theta)h)^n}{{r_{pcr}}^{n\delta}}\, ddr\, d\theta\, dx\, dh \\ &+\int_H f_h(h)\int_{r_{pcr}}^{\infty} f_{ccn}(x)\int_0^{2\pi}\int_R^L [P_{max}(x)g(\,d_{cp}(d,\theta)h]^n ddr\, d\theta\, dx\, dh\Bigg\} \end{aligned} \tag{21}$$

In the subsequent section in this paper, we show experimental results which show the effect of unknown primary receiver location on interference in comparison with a known primary receiver location. The experiment also shows that the challenges of the hidden primary user location also increase the interference in respect to higher mean and variance. It is also seen that the log-normal estimation satisfies both the derived CDF and Monte Carlo's simulations. However, we can also see the effect of some CR implementation parameters on the total interference under the power control scheme.

B. *Contention Control Scheme in when primary receiver location is unknown*

In modeling interference in the contention control scheme, can derive the n^{th} cumulant following the same process which was used in (21) and utilizing the characteristics function in (5) then the cumulant k_n of the total interference becomes;

$$k_n = \frac{\lambda\pi q_{mh}}{i^n}\int_H f_h(h)\left[-R^2(ipg(R)h)^n + n(iph)^n\int_0^{g(R)} t^{n-1-\frac{2}{\gamma}}dt\right] dh\ = \lambda\pi q_{mh}\left(\frac{n}{n-\frac{2}{\gamma}}g^{n-\frac{2}{\gamma}}(R) - R^2 g^n(R)\right)p^n\int_H f_h\,(h)h^n dh = \frac{2p^n\left(1-e^{-\lambda\pi d_{min}^2}\right)e^{n\mu+\frac{\sigma^2}{2}}}{(n\gamma-2)d_{min}^2 R^{n\gamma-2}} \tag{22}$$

Considering the system which is represented in figure 1, and the proposed contention control earlier in this work, then the characteristics function of the entire interference can then be denoted as;

$$\varphi_I(\omega) = \lim_{i\to\infty} exp\{q_{mh}\lambda\pi D_L\left(E\left(e^{i\omega pg(V)h}\right)-1\right)\}$$
$$= \lim_{i\to\infty} exp\left\{q_{mh}\lambda\pi D_L\left(\left(\int_H f_h(h)\int_0^{2\pi}\frac{1}{2\pi}\int_R^L exp\left[iwpg(d_{cp}(d,\theta)h\right]\frac{2d}{DL}dd\ d\theta\ dh-1\right)\right)\right\}$$
$$= \lim_{l\to\infty} exp\left\{q_{mh}\lambda\int_H f_h(h)\int_0^{2\pi}\int_0^{2\pi}\int_R^L\left[pg\left(d_{cp}(d,\theta)\right)h\right]^n r-dr\ d\theta\ dh\right\} \quad (23)$$

When we implement the log-normal estimation as used in the power control scheme, we can then derive the k^{th} cumulant of the interference as follows;

$$k_n$$
$$= \lim_{i\to\infty}\lambda\left\{\int_H f_h(h)\int_0^{r_{pcr}} f_{ccn}(x)\int_0^{2\pi}\int_R^L\frac{\left(d^\delta P_{max}(x)g\left(d_{cp}(d,\theta)\right)h\right)^n}{r_{pcr^{n\delta}}}\ rdr\ d\theta\ dxdh\right.$$
$$\left.+\int_H f_h(h)\int_{r_{pcr}}^{\infty} f_{ccn}(x)\int_0^{2\pi}\left[P_{max}(x)g(d_{cp}\ (d,\theta))h\right]^n rdr\ d\theta\ dx\ dh\right\} \quad (24)$$

V. Results and Discussions

In this section we have shown experimental results of the total interference power from all CR transmitters utilizing the power control and contention control schemes. In this work, we have investigated how the unknown location of the primary receiver affects the interference in both the power control and contention schemes. Also from this work it has be proven that the proposed schemes is an efficient way of increasing CR-primary interference. Furthermore we have also investigated the effects of shadow fading on the aggregate interference on a CR networks which implements both interference management schemes.

In figure 2, we have shown the effect of unknown location of the primary user location affects interference. The figure shows the log-normal approximation compatibility with the derived CDF and Monte Carlo's simulation. The parameters used are defined as follows; R is the radius of the interference region, the density of the stationary Poisson point process is denoted as λ**.** The pathloss exponent is represented as γ, the power control range is denoted as r_{pcr}, the power control exponent is represented as δ the maximum power and distance between transmitters are represented by P_{max} and d_p respectively.

The following values for the parameters where used under the control scheme in figure 2; R= 200m, λ=3 user/10^4m^2, γ=4, r_{pcr}=20m, δ=4, P_{max}=1W, d_p=0.5R.

In figure 2, we have analyzed the impact of the unknown location of the primary user on the resulting interference where we assumed a pathloss only channel. It has been shown from the figure that the hidden primary user problem increases the interference in terms of mean and variance when compared to the scenario when there is perfect knowledge of the location of the primary user. Also the shows that the log- normal approximation is appropriate for both derived CDF and Monte Carlo's Simulation.

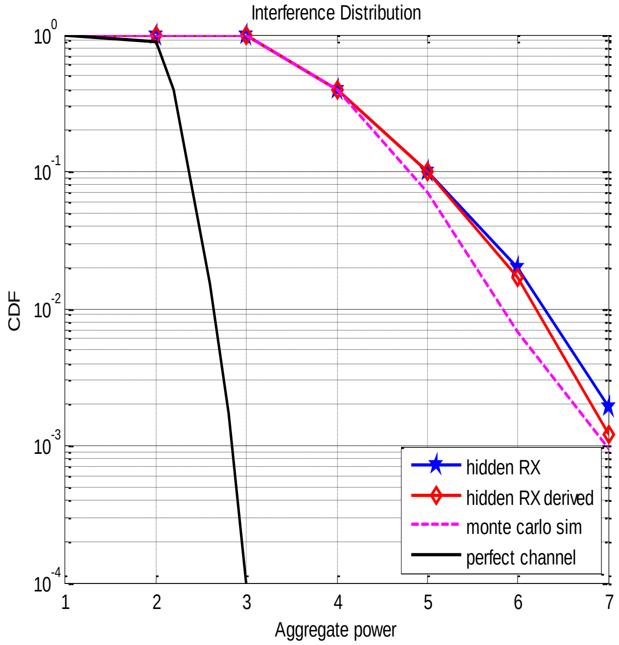


Fig. 2. Log normal approximation for interference distribution with hidden primary user location under Power control

In figure 3, we have shown the experimental results of the effect of the unknown location of the primary receiver under the contention control scheme. We setup the system just like the power control scheme with addition of the minimum contention distance between two CR transmitters d_{min}. We assumed a pathloss only channel, in regards to the presented results in figure 3; we can see that the uncertainty of the location of the primary user increases interference in terms of mean and variance [33]. Also the log – normal approximation for the interference is prone to inaccuracy [34] as the interferences increases when compared to the power control scheme in figure 2.

The following values for the parameters where used under the contention control scheme in figure 2; R= 100m, λ=3 user/10^{4}m^2, γ=4, d_{min}=20m, δ=4, P_{max}=1W, d_p=0.5R.

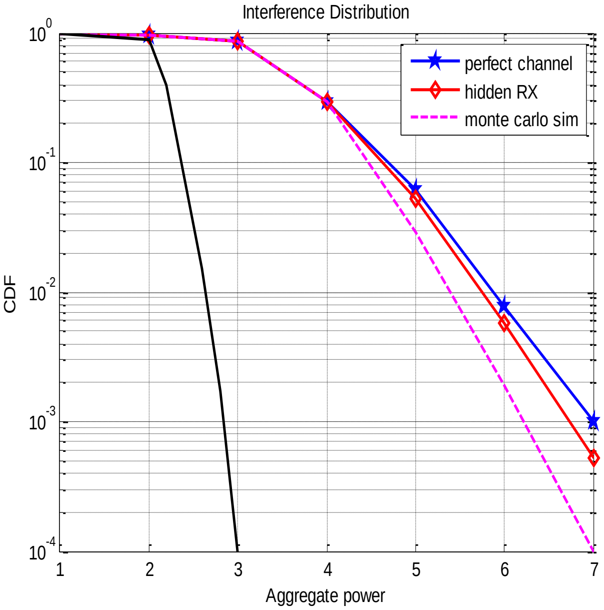


Fig. 3. Log normal approximation for interference distribution with hidden primary user location under Contention control

In this section, we have also investigated the impacts which the shadow fading has on the aggregate interference for CR network. The investigation was done considering various values of the Nakagami shape factor m under both schemes. The setup is the same as the initial setup as previously used with the additional standard variance σ_Ω=4dB.

The following values for the parameters where used under the Power control scheme; R= 100m, λ=3 user/10^{4}m^2, γ=4, r_{pcr}=20m, δ=4, P_{max}=1W, σ_Ω=4dB.

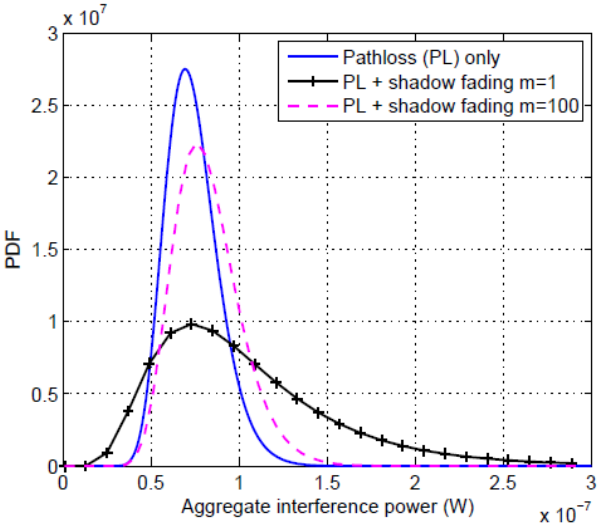


Fig. 4. The effect of shadow fading on the total interference power CR network under power control scheme

From figure 4, it can be seen that when the Nakagam-m=1, the interference channel becomes a Rayleigh channel which is influenced by the log-normal shadowing, however, when m=100, the instability or variations of the channel are greatly decreased. It is observed from figure 4, that the interference distribution possesses higher variance and heavier tails when shadow fading in integrated into the power control scheme.

The figure 5 shows the impact of shadow fading on the interference distribution under the contention control scheme. The setup is exactly like the power control with the exception of the power control range (r_{pcr}) which is replaced withd_{min}.

The following values for the parameters where used under the Power control scheme; R= 100m, λ=3 user/10^{4}m^2, γ=4, d_{min}=20m, δ=4, p=1W, σ_Ω=4dB

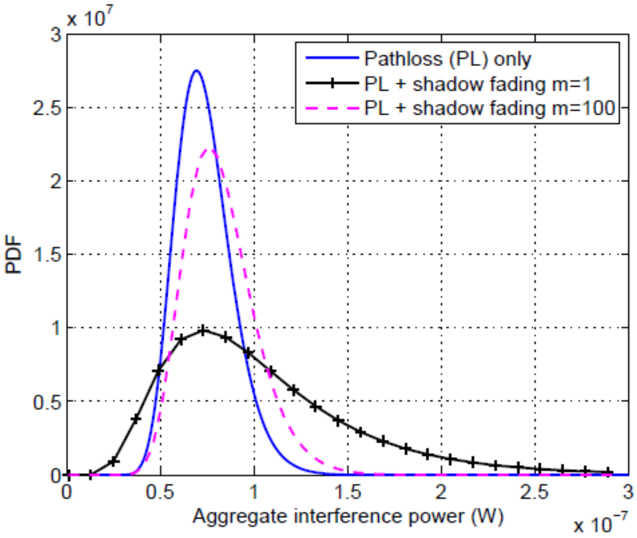


Fig. 5. The effect of shadow fading on the total interference power CR network under Contention control scheme

The incorporated shadow fading in the contention control scheme has similar effect with that of the power control. Hence it produces similar result.

Furthermore when exploiting a CR network under the power control scheme, its resulting interference can be regulated by altering the parameters which include P_{max} , r_{pcr} , λ and R. Figure 6 shows the effect of different CR implementation on the aggregate interference in a CR network, it has been shown that the interference can be reduced by either reducing the maximum transmission power P_{max} and or CR density λ or increasing power control range r_{pcr} and or the interference radius. However, it has been shown that modifying the IR radius is a good method to manage the interference.

This is due to the fact that the interference has high sensitivity to the IR radius than any other parameter as shown in figure 6. The previous setup for the power control scheme was retained with the exception of manipulating the parameter P_{max} , r_{pcr} , λ and R.

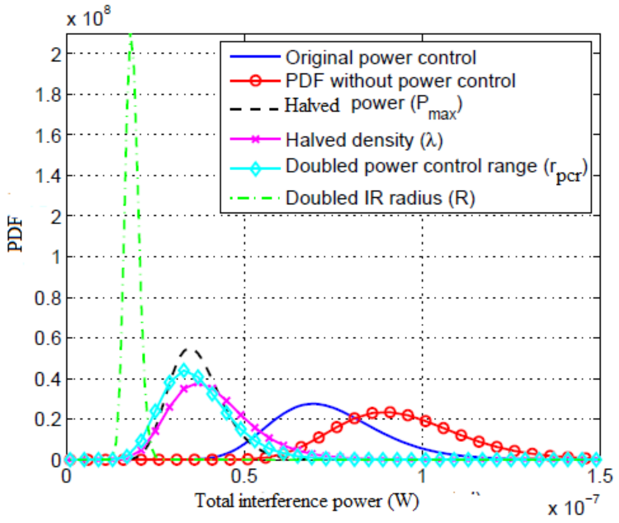


Fig. 6. Effect of different CR implementation on the total interference for CR network under power control

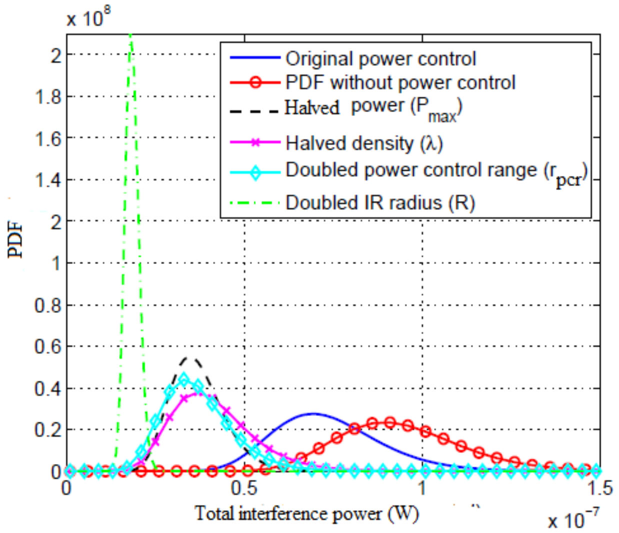


Fig. 7. Effect of different CR implementation on the total interference for CR network under Contention control

Figure 7 shows the impact which the different CR implementation has on the total interference for CR networks under the contention control scheme. It can be seen from figure 7, that the resulting interference has reduced mean just like that in the power control scheme, however the interference is reduced by decreasing p, λ and/ or increasing R or d_{min}. When comparing figure 6 and 7, it can be seen that increasing the interference radius is a good method in reducing the interference for both power and contention control scheme. Although, the power control scheme has higher sensitivity to the interference radius than the contention scheme. It also shows that when the transmission power or the CR transmitter density is reduced, the effect on the interference is the same in both schemes.

VI. Conclusion

In this work, we have investigated the interference at the primary receiver generated by the CR transmitters using power and contention control schemes. The power and contention control scheme have been estimated analytically, the interference distribution for the power and contention control have been estimated by log-normal distribution utilizing the cumulant based method. Both schemes have been proven to be good candidates in reducing interference at the primary receiver which is generated by the CR transmitters. The impacts of the unknown location of the primary receiver on the CR-primary interference was also investigated and it was found out that the unknown location of the primary receiver leads to higher CR-primary interference. Finally, it has been shown numerically that the impacts of the different CR implementation parameters on the resulting total interference under the power and contention control scheme.

APPENDIX A: Derivation of Equation (5)

Using same steps which were used in [25], we can the express the characteristics function of the total interference as follows;

$$\varphi_I(\omega) = \lim_{i\to\infty} e^{\lambda\pi\left(L^2-R^2\right)(Q-1)} \qquad \text{A.1}$$

where

$$Q = E\left(e^{i\omega Pg(V)H}\right)$$

$$= \int_H f_h\,(h) \int_R^L E\left[e^{i\omega Pg(r)h}\right]\frac{2r}{L^2-R^2}\,dr\,dh$$

$$= \int_H f_h(h) \int_R^L\left[(1-q_{mh}) + q_{mh}e^{i\omega pg(r)h}\right]\frac{2r}{L^2-R^2}\,dr\,dh \qquad \text{A.2}$$

The integral in (A.1) can be rewritten as;

$$\lim_{i\to\infty} \int_H f_h(h) \int_R^L e^{i\omega pg(r)h}\,\frac{2r}{L^2-R^2}\,dr\,dh =$$

$$1 + \frac{1}{L^2-R^2}\int_H f_h(h)T(\omega ph)dh$$

A.3

with $T(\omega ph)$given then we substitute (A.2) and (A.3) into (A.1) to obtain (5)

APPENDIX B: Derivation of Equation (10)

Substituting (3) and (4) into (11) we then have the following;

$$\varphi_Y(\omega) = exp\left\{\lambda\pi\int_H f_h\,(h)\int_{r_{ccn}} f_{ccn}(r)\left[R^2\left(1-e^{i\omega g(R)p(r)h}\right)+\right.\right.$$

$$\left.\left.i\omega P_{pcr}(r_{ccn})^h\int_0^{g(R)}(g^{-1}(t))^2e^{i\omega tp(r)h}dt\right]dr\,dh\right\} =$$

$$exp\left\{\lambda\pi\int_H f_h\,(h)\int_0^{r_{pcr}} f_{ccn}(r)\left[R^2\left(1-e^{i\omega(\frac{r}{r_{pcr}})^\delta P_{\max}g(R)h}\right)+\right.\right.$$

$$\left.\frac{i\omega r^\delta P_{max}h}{r_{pcr}^\delta}\int_0^{g(R)}(g^{-1}(t))^2e^{i\omega t(\frac{r}{r_{pcr}})^\delta P_{max}\,h}dt\right]drdh +$$

$$\lambda\pi\int_H f_h(h)\int_{r_{pcr}}^\infty f_{ccn}(r)\left[R^2\left(1-e^{i\omega g(R)P_{max}\,h}\right)+\right.$$

$$\left.\left.i\omega P_{max}h\int_0^{g(R)}(g^{-1}(t))^2e^{i\omega tP_{max}h}dt\right]dr\,dh\right\} \qquad \text{B.1}$$

The characteristic function in (10) is obtained by using (9) and (B.1)

APPENDIX C: Derivation of (14)

In a situation where the first equality of (C.1) holds according to [15]. Then equation (10) is derived instantly from (C.1).

$$K = \int_H f_h\,(h) \int_p f_p\,(p)\sqrt{hp\ dp\ dh} = \int_H f_h\,(h)\sqrt{h\ dh}$$
$$= \int_p f_p\,(p)\sqrt{p}\ dp$$
$$= \sqrt{P_{max}}\ \int_H f_h\,(h)\sqrt{h}\ dh\ (\int_0^c 2\pi r\lambda e^{-\lambda\pi r^2}(\frac{r}{c})^{\frac{\delta}{2}}\ dr$$
$$+ \int_c^\infty 2\pi\lambda r e^{-\lambda\pi r^2} dr) = \sqrt{P_{max}} \int_H f_h\,(h)\sqrt{h\ dh}$$
$$\left(\int_0^c 2\pi r\lambda e^{-\lambda\pi r^2}\ \left(\frac{\delta}{c}\right)^{\frac{\delta}{2}} dr + e^{-\lambda\pi c^2}\right) \qquad \text{C.1}$$

APPENDIX D: Derivation of (15)

$$\varphi_I(\omega) = \lim_{l\to\infty} exp\{\lambda\pi Di\big(E\big(e^{i\omega P_{pcr}g(V)h}\big) - 1\big)\} =$$
$$\lim_{i\to\infty} exp \left\{ \begin{matrix} \lambda\pi Di[\int_H f_h(h) \int_0^\infty f_{ccn}(x) \int_0^{2\pi} \frac{1}{2\pi} \int_R^L \exp[i\omega P_{pcr}(x) g(d_{cp} \\ (d,\theta))h] \frac{2r}{Dl}\ dr\ d\theta\ dx\ dh - 1] \end{matrix} \right\} =$$
$$\lim_{i\to\infty} exp \left\{ \begin{matrix} \lambda \int_H f_h\,(h) \int_0^\infty f_{ccn}(x) \int_0^{2\pi} \exp\left[\omega P_{pcr}(x) g\left(d_{cp}(d,\theta)\right) h\right] \\ r - rdr\ d\theta\ dx\ dh \end{matrix} \right\}$$

D.1

where $Dl = l^2 - R^2$ the first equality in (D.1) is derived in similar way as (A.1) and (A.2) then (15) is derived immediately from (D.1)

References

[1] FCC, “Spectrum policy task force report”, Technical report 02-135, Federal Communication Commission, Nov.2005.

[2] Y. Zhang, “Dynamic spectrum access in Cognitive radio wireless networks”, Proceedings of IEEE International conference on communication (ICC) 19-23, Beijing, China, May. 2008, pp.4927-4932.

[3] B.Wang and K.J.R.Liu, “ Advances in cognitive radio networks: A Survey,” IEEE Journal of selected topics in signal processing, vol.5, no.1, pp.5-22, Feb. 2011.

[4] S.Haykin. “Cognitive Radio: Brain-Empowered wireless commumcation”, IEEE Journal on selected areas in communication vol.23, no.2. pp.201-220, Feb.2005.

[5] I.F. Akyilidiz, W.Y.Lee and K..R. Chowdhury,” Spectrum management in cognitive radio ad hoc network,” IEEE networks, vol 23, no.4,pp.6-12, Aug.2009.

[6] C.X.Wan, X.Hong, H.H.Chen and J.Thompson, “On capacity of cognitive Radio Network with average interference power constraints,” IEEE Transaction on wireless communication, vol.8.no.4, pp. 1620-1625, April 2009.

[7] IEEE P802.22, fundamental requirements for the 802.22 WRAN standard, IEEE 802.22-05/0007r48, 29 Nov. 2006.

[8] S.A. Jafar and Srinivasa, “Capacity limits of cognitive radio with distributed and dynamic spectral activity”, in proceeding of IEEE international conference on communication (ICC), Istanbul, Turkey, June 2006, pp.5742-5747.

[9] Y.Y.Mihov, “Cross-layer Analysis and performance evaluation of cognitive radio networks,” The sixth international conference on systems and networks communications, (ICSNC) 23-29, Oct, Barcelona, Spain, pp.99-104.

[10] Z.Chen, C.X.Wong, X.Hong, J.Thompson, S.A.Vorobyov, X.Ge, H.Xiao and F.Zhao, “Aggregate Interference modeling in cognitive radio network with power and contention control,” IEEE Transaction on communication, vol.60, no.2, pp.456-468, Feb.2012.

[11] A.S.Hosseinzadeh-Salati, M.Nasiri-Kenari, “Aggregate interference modelling and static resource allocation in closed and open access femtocells,” IET communication, vol.8, no.7,pp.1007-1016.

[12] X.Yang and A.P.Pertropulu, “Co-channel interference modeling and analysis in a poisson field of interfers,” in proceeding of IEEE 40th Annual conference on information sciences and systems, Princeton, USA, Mar.2006, pp.432-437.

[13] X.Hong, C.X.Wong, J. Thompson and Y.Zhang, “Demystifying white spaces,” in proceedings of IEEE International conference , circuit and systems, (ICCCAS), Xiamen, China, 25-27 May 2008, pp.350-354.

[14] C.Wijenayake, A.Madanayake, J.Kota and L.Bruton, “Space-Time spectra white spaces in cognitive radio: Theory, Algorithms and circuits,” IEEE Journal on emerging and selected topics in circuits and systems, vol.3, no.4, pp.640-653.

[15] S.Geirhofer and L.Tong, “ Dynamic spectrum access in the time domain: modeling and exploiting white space,” IEEE communication magazine, vol.45, no.5, pp.66-72, May 2007.

[16] M.Hoyhtya, T.Chen and A.Mammela, “Interference management in frequency, time and space domains for cognitive Radios”, in proceeding of wireless telecommunication symposium, 22-24, Apr.2009, pp.1-7.

[17] Y.M.Shobowale, K.A.Hamdi, “A unified model for interference analysis in unlicensed frequency bands,” IEEE Transactions on wireless communication, vol.8, pp.4004-4013, Aug.2009.

[18] M. Timmers, S.Pollen, A.Dejonghe, A.Bahai, L.Van der Pere and F.Catthoor, “ Accuumulative interference modelling for cognitive radio with distributed channel access”Third international conference on cognitive radio oriented wireless networks and communications (CROWNCOM), Singapore, 15-17 May 2008, pp.1-7.

[19] P.Lin, J.Jia, Q.Zhang and M.Hamdi, “Dynamic spectrum sharing with multiple primary and secondary users”, IEEE transactions on vehicular technology, vol.60, no.4, May 2011.

[20] D.Treeumnuk, D.C.Popescu, “Enhanced spectrum utilisation in dynamic cognitive radio with adaptive sensing”, IET Signal processing, vol.8, no.4, pp.339-346, Jun. 2014.

[21] N.Hoven, A.Sahai, “Power scaling cognitive radio,” International conference on wireless networks, communications and mobile computing, Maui, HI. USA, 13-16 Jun.2005, pp.250-255.

[22] R.Menon, R.M.Buehrer and J.Reed, “Outage probability based comparison of underlay and overlay spectrum sharing techniques”, in proceeding of IEEE Dynamic spectrum access networks, Baltimore, USA, 8-11 Nov. 2005, pp.101-109.

[23] R.Menon, R.Buehrer and J.Reed, “On the impact of dynamic spectrum sharing techniques on legacy radio system,” IEEE Transaction on wireless communication, vol.7,no.1, pp.4198-4207, Nov.2008

[24] I.E.Igbinosa, O.O.Oyerinde, V.M.Srivastava, S.H.Mneney, “Analysis of impact of hidden primary receiver on interference modeling in power, contention and hybrid control schemes” in proceeding of international conference on emerging trends in network and computer communication (ETNCC),17-20 May,2015, Windhoek, Namibia.in Press.

[25] A.Goldsmith, Wireless Communication, Cambridge; Cambridge university press, 2005.

[26] E.S.Sousa and J.A.Silvester, “Optimum transmission range in a direct-sequence spread-spectrum multihop pack radio network,” IEEE journal of selected areas on communication, vol.8, no.5, pp.762-771, Jun.1990.

[27] X.Hong, C.X.Wang and J.S.Thompson, “Interference modeling of cognitive radio networks,” in proceedings of IEEE vehicular technology conference, Singapore, 11-14 May. 2008, pp.1851-1855.

[28] D.Stoyan, W.S.Kendall and J.Meeke, Stochastic geometry and its applications, Chichester; John Wiley and Sons, 1986.

[29] H.Q.Nguyen, F.Bacceelli and D.Kofman, “A Stochastic geometry analysis of dense IEEE802.11 network,” in proceedings of IEEE INFOCOM, 6-12 May 2007, pp.1199-1207.

[30] I.E.Igbinosa, O.O.Oyerinde, V.M.Srivastava, S.H.Mneney, “Numerical analysis of the impact of shadow fading in power control and contention control interference management schemes in cognitive radio,” in proceedings of IEEE international conference on emerging trends in computer network and communication (ETNCC), 17-20 May 2015, Windhoek, Namibia, pp. 144 -147.

[31] M.Pratesi, F.Santucci and F.Graziosi, “Generalized moment matching for the linear combination of log-normal RVs; Application to outage

analysis in wireless systems," IEEE transaction on wireless communications, vol.5, no.5, pp.1122-1132, May 2008.

[32] R.Menon, R.M.Buehrer, J.H.Reed, "Impact of exclusion region and spreading in spectrum-sharing Ad hoc networking", in proceedings of 1st international workshop on technology and policy for accessing spectrum, 1-5 Aug. 2006, Boston, USA.

[33] C.C.Chan and S.V.Hanly, " Calculating the outage probability in a CDMA network with spatial poisson traffic," IEEE transaction on vehicular technology, vol.50, no.1, pp. 183-204, Jan.2001.2

[34] M.F.Hanif, P.J.Smith, P.A.Dmochowski, "Statistical interference modelling and deployment issues for cognitive radio systems in shadow fading environments", IET communications, vol.6, no.13, pp.1920-1929.

Design of ANFIS Estimator of Permanent Magnet Brushless DC Motor Position for PV Pumping System

TERKI Amel
Genie Electric department
Biskra University
Biskra, Algeria

MOUSSI Ammar
Genie Electric department
Biskra University
Biskra, Algeria

TERKI Nadjiba
Genie electric department
Biskra University
Biskra, Algeria

***Abstract*—This paper presents a new scheme for PMBLDC (permanent magnet brushless direct current) rotor position estimation based an ANFIS (adaptive network fuzzy inference system) estimator. The operation of such motor requires accurate rotor position knowledge. However, most of rotor position sensors produce undesirable effects such as mechanical losses and have other disadvantages. In order to overcome the disadvantages, sensorless scheme seems to offer great advantages. This work present an ANFIS estimator design. Combining the adaptive capability of the neural network to gather with the reasoning ability of the fuzzy logic in ANFIS modeling results in a fast responding and flexible model. This procedure lends itself perfectly adapted for complex system such as PV pumping systems.**

***Keywords*—Photovoltaic system; Brushless DC motor; ANFIS estimator; Speed controller**

I. Introduction

In early1980s, DC motors are widely used for PV pumping applications either with or without intermediate converters [1]. Starting, steady state, and transient performances of solar power fed DC motors for linear and centrifugal pump load torques were analyzed and showed a perfect matching [2]. However these motors a bulky, and require frequent maintenance. Nowadays solar power fed permanent magnet brushless DC (PMBLDC) motors were being used instead [2].

PMBLDC machine is more popular due its simple structure and low cost [1, 3]. These machines have the advantages of light weight, small size, simple mechanical construction, easy maintenance, good reliability, and high efficiency [1, 3, 4, 5]. In general, the motors are equipped with mechanical position detecting devices to provide proper commutation for power devices in the bridge inverter [6]. These mechanical devices, such as Hall Effect sensor, optical or inductive sensor produce undesirable effects such as mechanical losses and also have other disadvantages like change in mechanical design and requirement of maintenance.

In order to overcome these drawback, there is a need to develop a sensorless scheme for estimating rotor position.

In recent years, many sensorless drive methods have been proposed. Of these, the most popular one in the back emf based Filter (EKF) is used [7]. This method provides excellent speed response but requires heavy online matrix computing. An offline FEM assisted position and speed observer has also been studied in the literature [8]. Zero crossing of line to line PM flux linkage is used for estimating the speed and position [9]. Flux Linkage Observer (FLO) based on the integration of back EMF, with a simple start-up method is proposed [10].

In this paper the design of an ANFIS estimator of permanent magnet brushless DC motor in PV pumping system is presented. The Adaptive Neuro Fuzzy Inference System (ANFIS), is a neural network that is functionally the same as a Takagi–Sugeno type inference model. The ANFIS is a hybrid intelligent system that takes advantages of both ANN and fuzzy logic theory in a single system, by using the ANN technique to update the Takagi–Sugeno type inference model parameters. In order to explain the concept of ANFIS structure, five distinct layers are used. The first layer in the ANFIS structure is the fuzzification layer; the second layer performs the rule base layer; the third layer performs the normalisation of membership functions (MFs); the fourth and fifth layers are the defuzzification and summation layer respectively [11].

This rest of the paper is organized as follows. In section 2, the configuration of PV pumping systems is presented, The Adaptive network fuzzy inference system estimator is described in section 3. Section 4 shows the experimental performance of Adaptive network fuzzy inference system estimator. Finally, Section 5 concludes our contribution and merits of this work.

II. General System Layout

The full system mainly consists of the solar cell array generator, DC/DC converter with MPPT (maximum power point tracker) command, PMBLDC motor with its bridge inverter coupled to a centrifugal pump load. The motor is controlled through a hysteresis current loop and an outer speed with PI type controller as shown in fig1.

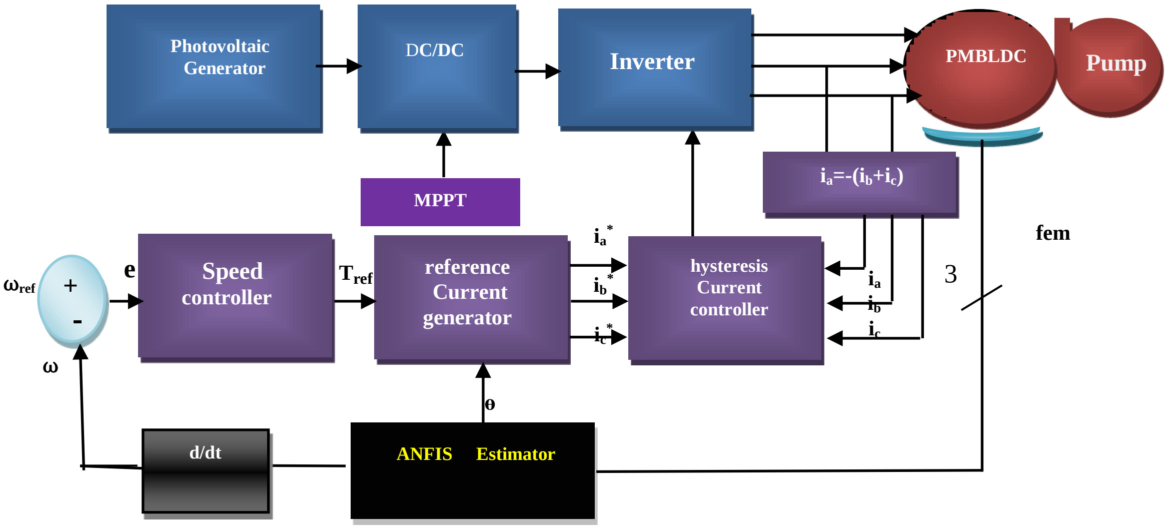


Fig. 1. Overall system configuration

A. PV generator model

The characteristic of the Photovoltaic generator can be presented by the following nonlinear equation [4]:

$$I = I_{sc} - I_o[\exp(\frac{(V+R_sI)}{V_{th}})-1]-\frac{(V+R_sI)}{R_{sh}} \quad (1)$$

Where:

I_{PV} Array output current,

R_{sh} PV array equivalent shunt resistance,

I_{sc} PV array short circuit current,

I_o PV array reverse saturation current,

R_s PV array series resistance,

V_{th} PV array thermal voltage.

The thermal voltage V_{th} and t reverse saturation current Io are successively identified by [4]:

$$V_{th} = \frac{(V_{op}+R_sI_{op}-V_{oc})}{\log(1-\frac{I_{op}}{I_{sc}})} \quad (2)$$

$$I_o = (I_{sc}-I_{oP})\exp\left(-\frac{(V_{op}+R_sI_{op})}{V_{rh}}\right) \quad (3)$$

B. Permanent magnet Brush-Less DC motor model

The simplified schematic of PMBLDC motor who has a trapezoidal electromotive force, the use of Park transform is not the best approach in modelling the machine. Instead the natural approach in used where the emf is generated with respect to rotor position [4]. The operating sequences of the machine can be subdivided into six cycles with respect to rotor position as shown in table II.

The electric of the motor can be described by [4]

$$V_{an} = Ri_a + p\lambda_a + e_a \quad (4)$$

$$V_{bn} = Ri_b + p\lambda_b + e_b \quad (5)$$

$$V_{cn} = Ri_c + p\lambda_c + e_c \quad (6)$$

With

$$V_{an} = V_{a0} - V_{n0} \quad (7)$$

$$V_{bn} = V_{b0} - V_{n0} \quad (8)$$

$$V_{cn} = V_{c0} + V_{n0} \quad (9)$$

Where R: per phase stator resistance. $i_{a,b,c}$ and $\lambda_{a,b,c}$ are respectively phase currents of phases a, b and c and total flux linkage of a,b and c. p: Laplace operator.

The flux expressions are given by the following expressions:

$$\lambda_a = L_S i_a - M(i_b + i_c) \quad (10)$$
$$\lambda_b = L_S i_b - M(i_a + i_c) \quad (11)$$
$$\lambda_c = L_S i_c - M(i_a + i_b) \quad (12)$$

Where Ls: the self-inductance and M: the mutual inductance.

And

$$i_a + i_b + i_c = 0 \quad (13)$$

Therefore by substituting Eq.13 in Eqs 10, 11 and 12:

$$\lambda_a = i_a(L_S + M) \quad (14)$$
$$\lambda_b = i_b(L_S + M) \quad (15)$$
$$\lambda_c = i_c(L_S + M) \quad (16)$$

From the electrical equations 4, 5 and 6, the following system is obtained

$$\begin{bmatrix} V_{an} \\ V_{bn} \\ V_{cn} \end{bmatrix} = \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + p \begin{bmatrix} L_{eq} & 0 & 0 \\ 0 & L_{eq} & 0 \\ 0 & 0 & L_{eq} \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \quad (17)$$

With Leq = L_S –M

From this system, the decoupled phase equations are obtained and the explicit current equations are given by:

$$p\begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} = \begin{bmatrix} {}^{1}/_{L_{eq}} & 0 & 0 \\ 0 & {}^{1}/_{L_{eq}} & 0 \\ 0 & 0 & {}^{1}/_{L_{eq}} \end{bmatrix} \left[\begin{bmatrix} V_{an} \\ V_{bn} \\ V_{cn} \end{bmatrix} - \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} - \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \right] \qquad (18)$$

The mechanical part is expressed by the following equation:

$$J\frac{d\Omega}{dt} + B\Omega = T_e - T_r \qquad (19)$$

With:

T_e: electromagnetic torque.

T_r: Load torque

Ω : speed

J : moment of inertia

B : viscose friction coefficient

Neglecting the frictional coefficient and taking $\Omega = \frac{\omega}{P}$ where P is the pole pairs number, (19) can written as:

$$\frac{d\omega}{dt} = P.(T_e - T_r)/J \qquad (20)$$

The developed torque can be expressed by

$$T_e = (e_a i_a + e_b i_b + e_c i_c)/\omega \qquad (21)$$

And the angular position is expressed by

$$\frac{d\theta}{dt} = \omega \qquad (22)$$

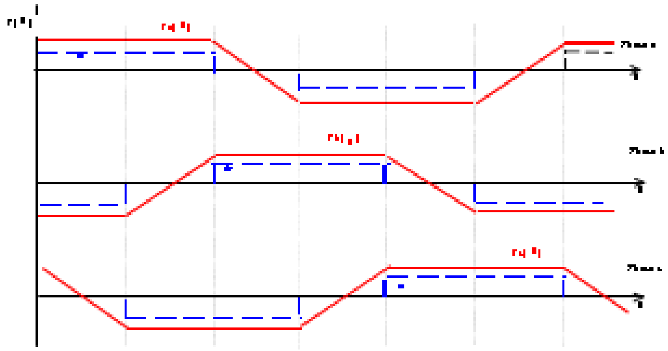


Fig. 2. Back e.m.f and current waves forms for phase a, b and c

C. Speed control

PI speed controller is widely used in industry due to its ease in design and simple structure. The rotor speed $\omega(k)$ is compared with the reference speed $\omega_{ref}(k)$ and the resulting error is estimated at the nth sampling instant as:

$$e(k) = \omega_{ref}(k) - \omega(k) \qquad (23)$$

$$\Delta e(k) = e(k) - e(k-1) \qquad (24)$$

The value of the torque reference is given by [4]:

$$T_{ref}(k) = T_{ref}(k-1) + K_P \Delta e(k) + K_i e(k) \qquad (25)$$

Where $e(k-1)$, the speed error of previous interval is, $e(k)$ is the speed error of working interval. k_P and k_i are speed controller gains.

D. Current control

Several techniques can be used to control the phase current of the PMBLDC motor. In this paper a hysteresis current controller is used. It has the major advantage of not requiring machine parameters to be known. However the commutation frequency is not constant [4]. It depends on many factors such as the applied voltage, the back emf, hysteresis band ΔI...etc.

Maximum value of commutation frequency is obtained at starting and is given by [4]:

$$F_{max} = U/8L_S\Delta I \qquad (26)$$

The commutations are obtained by comparing actual currents $i_{a.b.c}$ to a rectangular reference $i^*_{a.b.c}$ and by keeping them in hysteresis band ΔI. The commutation sequences of switches are summarised in the table I. [5].

TABLE I. THE CMMUTATION SEQUENCES OF SWITCHES

Si $i_a < (i_a^* - \Delta I)$	T_1 on	T_4 off	V_a= U/2
Si $i_a > (i_a^* + \Delta I)$	T_1 off	T_4 on	V_a= -U/2
Si $i_b < (i_b^* - \Delta I)$	T_2 on	T_5 off	V_b= U/2
Si $i_b > (i_b^* + \Delta I)$	T_2 off	T_5 on	V_b= -U/2
Si $i_c < (i_c^* - \Delta I)$	T_3 on	T_6 off	V_c= U/2
Si $i_c > (i_c^* + \Delta I)$	T_3 off	T_6 on	V_c= -U/2

E. Pump model

A centrifugal type is used. It can be described as an aerodynamic load which is characterised by the following load equation [4]:

$$T_1 = A.\omega^2 \qquad (27)$$

Where A is the pump constant

TABLE II. OPERATING SEQUENCES WITH RESPECT TO ROTOR POSITION

Rotor position $\theta°$	Phase voltages, back emf and reference current values								
	V_{ao}	e_a	i_a^*	V_{bo}	e_b	i_b^*	V_{co}	e_c	i_c^*
0°-60°	U/2	$K_e\omega$	I^*	-U/2	$-K_e\omega$	$-I^*$	0	$-6K_e\omega\theta/\pi + K_e\omega$	0
60°-120°	U/2	$K_e\omega$	I^*	0	$6K_e\omega\theta/\pi - 3K_e\omega$	0	-U/2	$-K_e\omega$	$-I^*$
120°-180°	0	$-6K_e\omega\theta/\pi + 5K_e\omega$	0	U/2	$K_e\omega$	I^*	- U/2	$-K_e\omega$	$-I^*$
180°-240°	U/2	$-K_e\omega$	$-I^*$	U/2	$K_e\omega$	I^*	0	$-6K_e\omega\theta/\pi + 7K_e\omega$	0
240°-300°	U/2	$-K_e\omega$	$-I^*$	0	$-6K_e\omega\theta/\pi + 9K_e\omega$	0	U/2	$K_e\omega$	I^*
300°-360°	0	$6K_e\omega\theta/\pi - 11K_e\omega$	0	U/2	$-K_e\omega$	$-I^*$	U/2	$K_e\omega$	I^*

III. ANFIS ESTIMATOR

The adaptive network fuzzy inference system play a significant role in the field of artificial intelligence. It combines the advantages of a fuzzy controller as well as quick response and adaptability nature of artificial neural network [11]. Hence using this scheme one can avoid the use of a position sensor which produces mechanical losses in the motor. Different configurations of adaptive network fuzzy inference system have been tested and the best one that gives the minimum error value is presented in fig6.

The Algorithm of the model structure was constructed as shown in Fig.3. The membership functions were obtained from the data set of the back emfs which were first normalized for loading Data. Then generate Fis, built on sugeno structure, which is obtained by creating of input membership function: number (tree inputs) & type (triangular) and output type(linear). The choice of optimization method (hybrid) is obtained by Train Fis. If the results are satisfactory (error is minimum) after test Fis, denormalized procedure is done and go to application thereafter. The ANFIS structure consists of 5 layers of neurons, each of which having a very specific behavior.

From these, second, third and fifth layer have constant behavior, while layers 1 and 4 have varying parameters.

Layer 1: consists of five adaptive neurons in which the fuzzification is performed, that is: the grade of membership to the defined membership functions of the input is evaluated.

$$f_i^1 = \mu_{Ai}(x) \tag{28}$$

Where x is the input to ith neurons and Ai is membership function correspond to variable x.

The membership function $\mu_{\mathrm{Ai}}(x)$ is triangular.

Layer2: consists of 125neurons, each node output represents the firing weight of a rule gives than:

$$W_k = \mu_{Ai}(x) \times \mu_{Bj}(y) \tag{29}$$

k : represent the number of rule, i : represent the number of x partition and j : number of y partition.

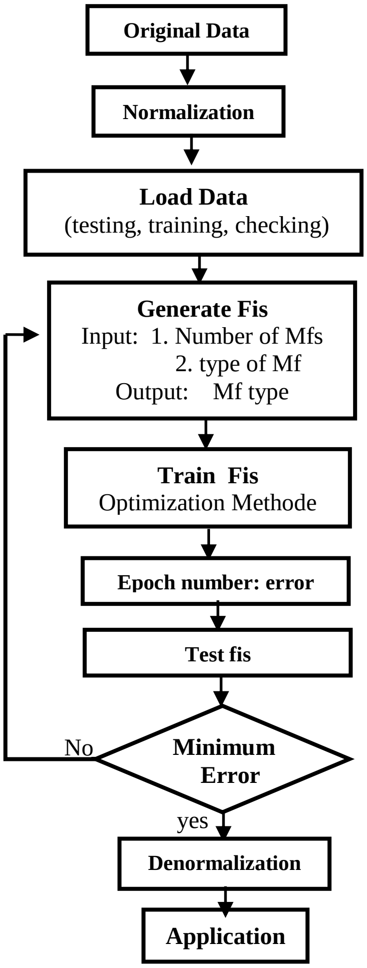


Fig. 3. Algorithm of the ANFIS model

Layer3: consists of 125 neurons, every node calculates the ratio of every rule firing weight to the sum of all rule firing weights.

$$\overline{W_k} = \frac{W_k}{\sum W_i} \tag{30}$$

Layer 4: consists of 125 neurons. Nodes in layer 4 are adaptive nodes in which the consequent evaluation inference is calculated, its output is defined as:

$$f_k^4 = \overline{W_k} \times f_k = \overline{W_k} \times (p_k x + q_k y + r_k) \quad (31)$$

$\overline{W_k}$ is the output of layer3 and $\{p_i, q_i, r_i\}$ is called consequent parameters.

Layer5: Finally, is single node in layer 5, where sums all the outputs from 4th layer to compute the overall output of the network:

$$f^5 == \sum_k \overline{W_k} \times f_k^4 \quad (32)$$

IV. Simulation Results

The entire ANFIS network architecture is represented in the Fig.6. The inputs are the three line to line back emfs of the motor namely Eab, Ebc and Eca Fig4. And the desired output is the estimated rotor position angle. For producing the above relationship with least possible error, a five layer feed forward adaptive network fuzzy inference system is used. The input layer consists of five neurons that have their inputs as the three line to line back emfs, the three hidden layers consists of 125 each and the output layer consists of a single neuron whose output is the estimated position of the rotor. The relationship between the back emf and the position of the rotor is shown in the Fig.5.

Fig.7 shows estimated rotor position by adaptative network fuzzy inference system and the error, along with real rotor position it’s a case without regulation. Fig.8 shows the estimated rotor position by adaptive network fuzzy inference system and the error, along with real rotor position the case is with PI speed control. In the comparison of the cases with control (PI regulator) and at without control, The controlled system attains the steady state in (0.04s) greater than the case without regulation the system attains the steady state in(0.023s), so the PI regulator slows the dynamic responses.

The performance of thThe performance of the trained adaptive network fuzzy inference system is found to have very minimal errors witch is about 1.303e^{-7} as seen in the Fig.7 (without control).

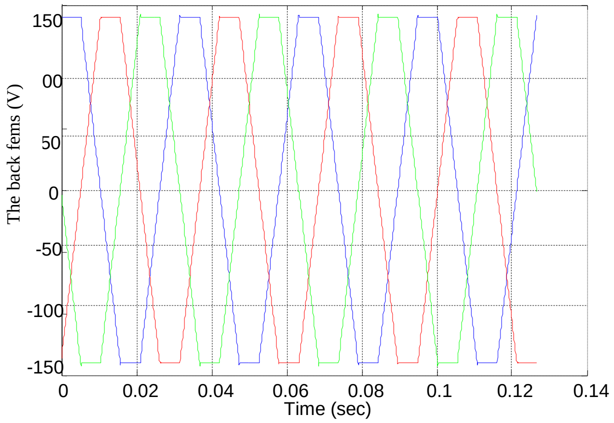


Fig. 4. The inputs curves

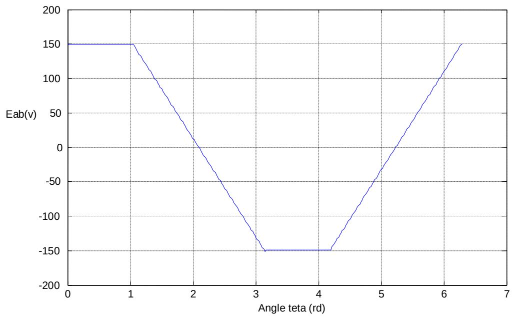


Fig. 5. The back emf Eab

For Fig.8 (with PI controller), the error is 7.8386e^{-9}. The error of output of ANFIS and real rotor position sensor are compared through Fig.7 and Fig.8 respectively. It is clear that the error in Fig.8 is too small (7.8386e^{-9}) than these in Fig.7 (1.303e^{-7}). The proposed scheme works efficiently where the usage of neural network topology together with fuzzy logic in The adaptive network, ANFIS, not only includes the characteristics of both methods, but also eliminates some disadvantages of their lonely-used case.

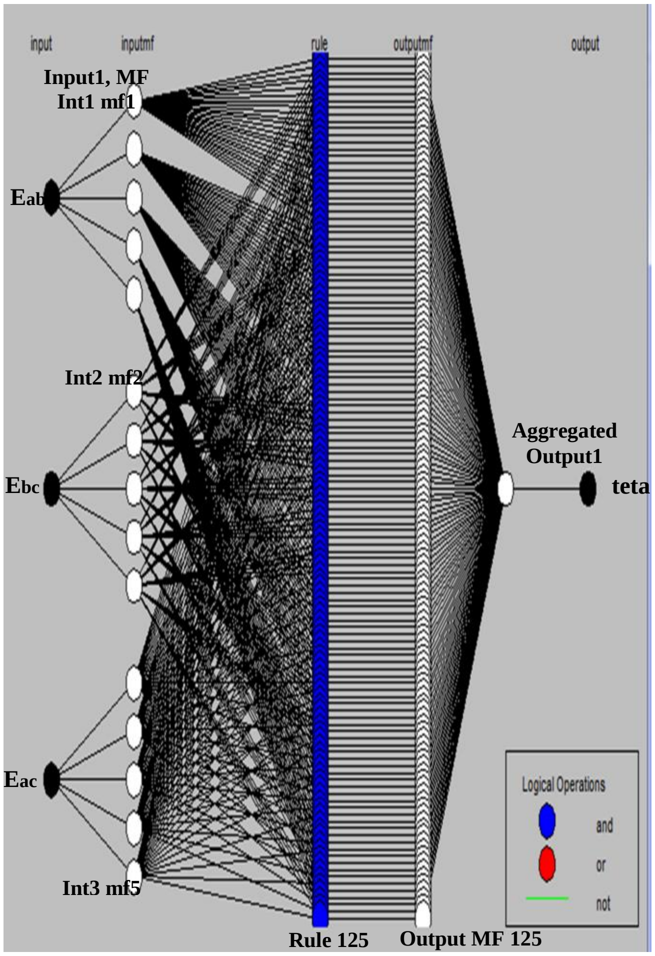


Fig. 6. Structure of ANFIS model proposed

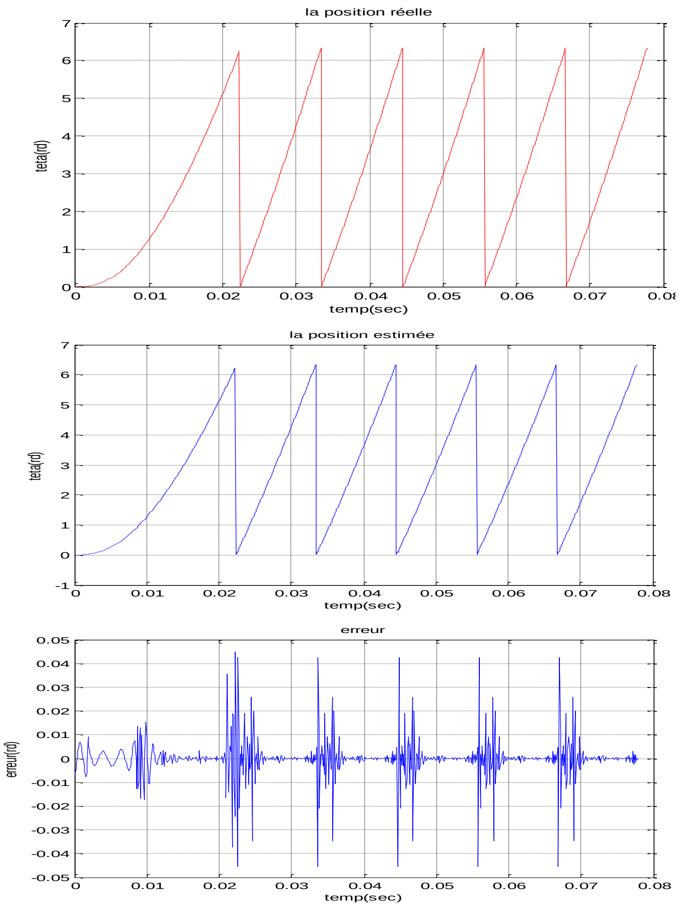
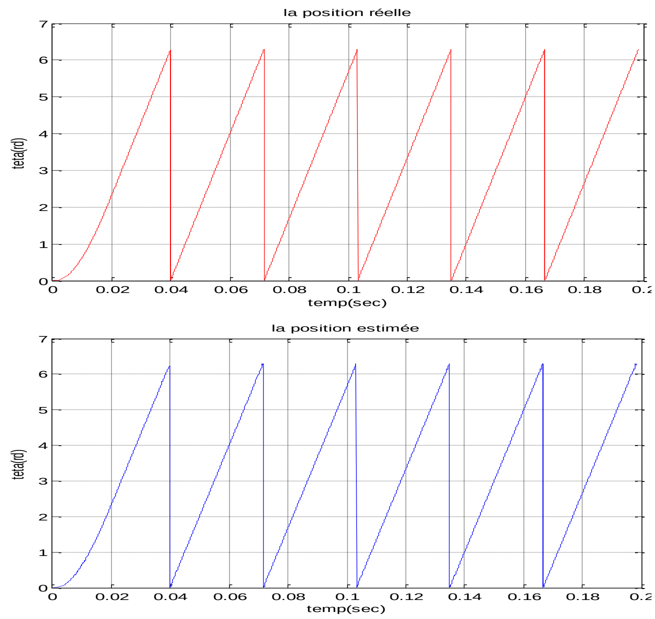


Fig. 7. Simulation Results without control



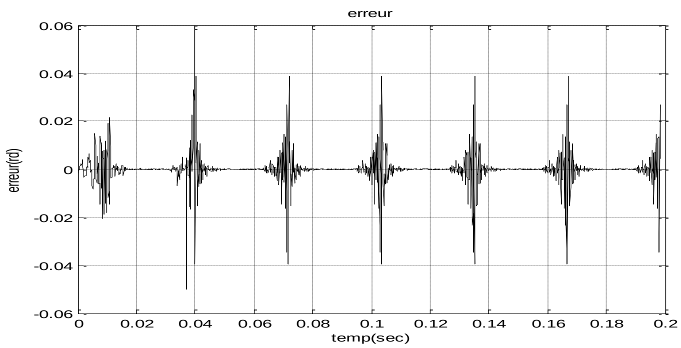


Fig. 8. Simulation Results with PI controller

V. Conclusion

The developed ANFIS models were successful in estimating rotor position of PMBLDC motor used in PV pumping system. In this paper the design of adaptive network fuzzy inference system based sensorless scheme was developed for permanent magnet brushless DC motor. The proposed scheme works efficiently for both with control (PI regulator) and at without control. It was found that the error is very minimal and this is smaller in system with regulation, than these without regulation. Also adaptive network fuzzy inference system based estimator is very flexible.

This scheme not only eliminates the position sensor, thereby cutting the cost but also transforms the drive into a highly efficient drive by eliminating losses caused by position sensors. The estimation error is found to be very minimal proving that the developed system is a very efficient and a reliable one.

References

[1] A.Moussi, A. Saadi, A. achour, Greg Asher, Photovoltaic pumping systems technologies trends, Larhyss journal ,2003,pp.127-150.

[2] PackiamPeriasamy, N.K.Jain,I.P.Singh, A review on development of photovoltaic water pumping system, Renew.and Sustainable.Energ. Reviews, Vol. 43 ,2015, pp. 918–925.

[3] MetinDemirtas,AslanDenizKaraoglan, Optimization of PI parameters for DSP-based permanent magnet brushless motor drive using response surface methodology, Energy Conversion and Management, Vol. 56, 2012, pp. 104–111.

[4] A. Terki, A. Moussi, A. Betka, N. Terki, An improved efficiency of fuzzy logic control of PMBLDCfor PV pumping system, Appl.Math.Model , Vol. 2, 2012, pp. 934-944.

[5] A.Moussi, A.Terki, Greg Asher, Hysteresis current controller of permanent Magnet Brushless DC motor PV pumping system, ASME, 2005, pp. 523-528 [International Solar Energy conference].

[6] T. Hiyama, Neural network based estimation of maximum power generation from PV modules using environmental information, IEEE Trans. EC, Vol.12, 199, pp. 241–247.

[7] A. Ungurean, V. Coroban-Schramel and I. Boldea, Sensorless control of a BLDC PM motor based on I-f starting and Back-EMF zero-crossing detection, OPTIM ’10, 2012, pp. 377-382 [Optimization of Electrical and Electronic Equipment Conference,OPTIM ’10.IEEE 12th International].

[8] AlinStirban, Ion Boldea, and Gheorghe-DanielAndreescu, Motion-Sensorless Control of BLDC-PM Motor With Offline FEM-Information-Assisted Position and Speed Observer, IEEE Trans. Ind. Appl, Vol. 48, 2012, pp. 1950-1958.

[9] LiviuIoanIepure, Ion Boldea and FredeBlaabjerg, Hybrid I-f starting and observer-based sensorless control of single phase BLDC-PM Motor drives, IEEE Trans. Ind. Electron, Vol.59, 2012,pp. 3436 – 3444.

[10] SreepriyaR&RagamRajagopal, Sensorless Control of Three Phase BLDC Motor Drive with Improved Flux observer, ICCC, 2013, pp.292-297 [IEEE International Conference on Control Communication and Computing].

[11] Ali M. Abdulshahed, AndrewP. Longstaff, Simon Fletcher, Alan Myers, Thermal error modelling of machine tools based on ANFIS withfuzzy c-means clustering using a thermal imaging camera, Appl. Math. Model, Vol. 39, 2015, pp. 1837–1852.

Appendix

The PV generator, motor and pump used in this study have the following parameters: PV generator Modules AEG-40.
(Temperature T=25°Cand solar insolation E=1000W/m².)

Open circuit voltage	22.40 V
Short circuit current	2.410 A
Series resistance	0.450 Ω
Current temperature coefficient	0.06%/ C
Voltage temperature coefficient	0.40%/ C
Centrifugal pump	
Rated speed	3000 rev/min
Rated power	521 W
Flowrate	2.597 l/s
Head	14.11 m
Efficiency	69%
Brushless DC motor	
Rated power	690 W
Rated speed	3000 rev/min
Rated voltage	200-220V
Rated current	4.8 A
Per phase resistance	1Ω
Per phase inductance	5 mH
Poles number	6
E.m.f constant	0.47

Effective Calibration and Evaluation of Multi-Camera Robotic Head

Petra Kocmanova
LTR s.r.o.
Brno, Czech Republic

Ludek Zalud
LTR s.r.o.
Brno, Czech Republic

***Abstract*—The paper deals with appropriate calibration of multispectral vision systems and evaluation of the calibration and data-fusion quality in real-world indoor and outdoor conditions. Checkerboard calibration pattern developed by our team for multispectral calibration of intrinsic and extrinsic parameters is described in detail. The circular object for multispectral fusion evaluation is described as well. The objects were used by our team for calibration and evaluation of advanced visual system of Orpheus-X3 robot that is taken as a demonstrator, but their use is much wider, and authors suggest to use them as testbed for visual measurement systems of mobile robots. To make the calibration easy and straightforward, the authors developed MultiSensCalib program in Matlab, containing all the described techniques. The software is provided as publicly available, including source code and testing images.**

Keywords—calibration; camera; mobile robot; thermal imaging; data-fusion

I. Introduction

Reconnaissance mobile robotics gains importance during the last years. Visual and space measurement subsystem is typically the most important sensory equipment with most significant impact on mission success. There are many missions in today’s society that may require expendable robots to perform exploration in inaccessible or dangerous environments instead of indispensable people. As examples we can name CBRNE (Chemical, biological, radio-logical, nuclear, explosive), counter-terrorist fight, US&R (Urban Search and Rescue), etc.

Since the missions take place in real world, the robots have to be equipped to most, if not all, possible conditions that may happen. During both military and non-military search and rescue missions, the robot can meet such conditions like complete darkness, smoke, fog, rain, etc. For these conditions, the visual spectrum of humans is not sufficient to provide valuable data. One of the most promising approaches for a wide spectrum of situations is a combination of data from the visual spectrum, near infrared spectrum and far infrared spectrum. In visual spectrum (using standard tricolor cameras), the operator has the best overview of the situation since he/she gets a signal that is most similar to what he knows. By using thermal imagers working in far infrared spectrum he/she can perfectly perceive even slight changes in temperatures. This spectrum very well penetrates through water particles (fog, rain) plus it is not affected by visible light conditions. Most TOF (time-of-flight) proximity scanners and cameras work in the near-infrared spectrum.

The main aim of this paper is the determination of effective sensory head calibration, containing typical sensors for the mentioned situations – tricolor cameras working in the visual spectrum, thermal imagers working in far infrared (FIR) and proximity camera working in near infrared (NIR).

Optimal image configuration is an important factor for effective calibration, so great attention was paid to it.

Calibration of the sensory head was proposed according to Zhang algorithm [1]. Zhang investigated the performance of his one camera calibration algorithm with respect to a number of images of the model plane. A number of images varied from two to 16. The error of intrinsic and extrinsic parameters decreased significantly between calibration from two and three images. Precision improves for more than images only insignificantly.

Calibration performance with respect to the orientation of the model plane was also investigated in [1]. Best performance was achieved with angle 45° between calibration plane and the image plane. This angle value is difficult to apply in real condition because it decreased the precision of corners extraction.

Photogrammetric software Photomodeler [2] recommends for one camera calibration using minimal six and optimal eight images of calibration plate from different angles. Another recommendation is using less than 12 images for camera lenses with wide angle and high distortion. Next recommendation is making at least two images with a roll of 90° (camera portrait, landscape orientation). Unfortunately this rotation isn’t possible with proposed sensory head, because of the sensory head manipulator.

Bouquet in Complete Camera Calibration Toolbox for Matlab [3] recommends using about 20 images of a planar checkerboard. 6 – 10 images should be enough for calibration in Omnidirectional Camera Calibration Toolbox for Matlab [4].

Effective image configuration for camera calibration of the sensor head of robot Orheus-X3 will be investigated in this paper.

The organization of the paper is as follows. In Chapter II the used hardware is described. Chapter III deals with calibration process of the camera head. In Chapter IV the data-fusion is described, and Chapter V aims to optimal image configuration and describes evaluation experiments made to evaluate the system.

II. Hardware

Although the CASSANDRA robotic system is rather complex and contains several interesting robots, only the Orpheus-X3 is important for the purposes of this paper. CASSANDRA robots are described in detail in [10], [11], [12].

A. *Orpheus-X3*

The Orpheus-X3 is an experimental reconnaissance robot based on the Orpheus-AC2 model made by our team to facilitate the measurement of chemical and biological contamination or radioactivity for military. The Orpheus-X3 offers the same drive configuration as its predecessor, namely the four extremely precise AC motors with harmonic gears directly mechanically coupled to the wheels; this configuration makes the robot very effective in hard terrain and enables it to achieve the maximum speed of 15 km/h. The main difference lies in the chassis, which is not designed as completely waterproof but consists of a series of aluminum plates mounted on a steel frame of welded L-profiles. This modular structural concept makes the robot markedly more versatile, and this is a very important aspect in a robot made primarily for research activities. Furthermore, the device is equipped with a 3DOF manipulator for the sensor head. The manipulator, again, comprises very powerful AC motors combined with extremely precise, low backlash harmonic drive gearboxes made by the Spinea company. The presence of such precise gearboxes can be substantiated by several reasons, mainly by the fact that the robot is used not only for telepresence but also for mobile mapping and SLAM. As currently planned, the robot's only proximity sensor is the TOF camera placed on the sensory head. The Orpheus robots are described in more details in our previous papers, such as [5].

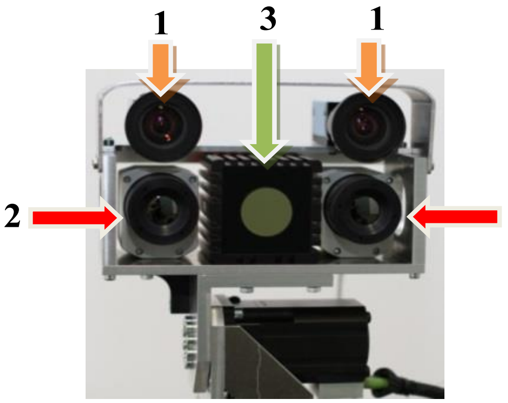


Fig. 1. The sensor head. 1 – tricolor CCD cameras; 2 – thermal imagers; 3 – TOF camera

B. *Sensor Head*

The sensor head containing five optical sensing elements is shown in Fig. 1. The sensors are as follows:

- Two identical tricolor CCD cameras (see 1 in Fig. 1): TheImagingSource DFK23G445 with the resolution of 1280x960 pixels, max. refresh rate of 30Hz, and GiGe Ethernet protocol. This device is equipped with a Computar 5mm 1:1.4 lens. The field of view is 40˚(h) x 51°(v).
- Two identical thermal imagers (see 2 in Fig. 1): Flir Tau 640 with the resolution 640x512, temperature resolution 0.05K and Ethernet output. The field of view is 56˚(h) x 69˚(v).
- One TOF camera (see 3 in Fig. 1): A Mesa Imaging SR4000 with the range of 10m, resolution of 176x144 pixels, and an Ethernet output. The field of view is 56˚(h) x 69˚(v).

The largest FOV capture thermal imagers and the TOF camera, which is required for the simultaneous use of stereovision and thermal stereovision. The main disadvantage of the applied TOF camera is its low number of pixels. Compared to the CCD cameras, it is about 10 times lower in one axis, and in relation to thermal imagers it is 4 times lower.

III. Sensor Head Calibration

Here will be described only calibration of intrinsic and extrinsic parameters. It is also necessary to calibrate temperatures of thermal imagers, in detail described in [6] and TOF camera measured distances, this calibration is described in detail in [7].

First condition for successful calibration is calibration plate with pattern visible in all 3 used spectrums:

- Infrared for TOF camera (850 nm).
- Visible spectrum for CCD cameras.
- Long-wavelength infrared for thermal imagers.

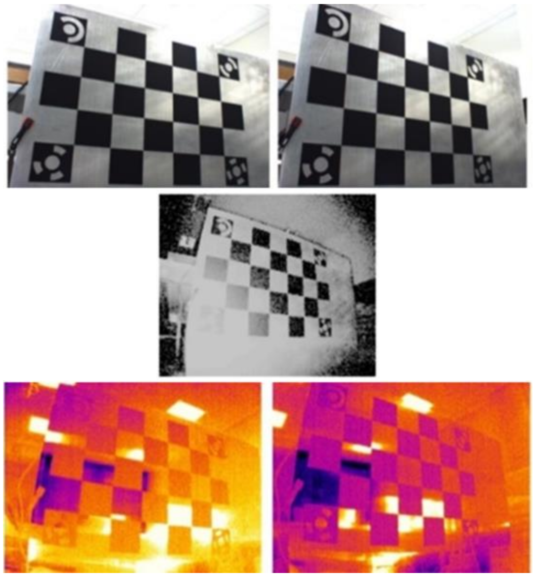


Fig. 2. The initial calibration plate: the left and right CCD cameras (up); the TOF camera intensity image (center). the left and right thermal imager

Three calibration plate based on checkerboard pattern were proposed. At first sufficient contrast of the calibration pattern should be achieved only by different materials. This version comprised an aluminum panel (low emissivity; high reflectivity) and a self-adhesive foil (high emissivity; low reflectivity). The main problem related to this initial board consisted in the high reflectivity of the aluminum base in cases that images are acquire under non-perpendicular angle (see Fig. 2).

The second version consisted of an aluminum panel with a laser-cut, anodized pattern and chipboard covered by a black, matt foil. Anodizing of aluminum panel reduces high reflectivity. Good contrast of checkerboard pattern for thermal imagers was achieved by heating of aluminum part at 50°C.

The final version included a 2 mm laser-cut aluminum plate with active heating. This version is more comfortable and shortens time needed to prepare calibration (see Fig. 3).

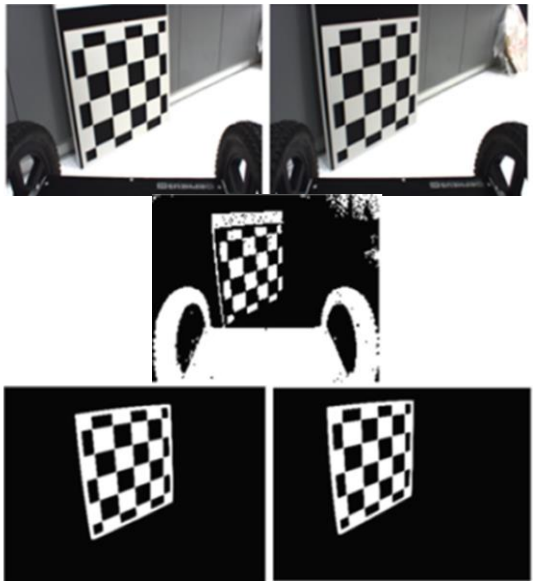


Fig. 3. The final calibration plate: the left and right CCD cameras (up); the TOF camera intensity image (center). the left and right thermal imager cameras (down)

Software *MultiSensCalib* was created for calibration of intrinsic and extrinsic parameters of sensor head camera system. The calibration comprises the following stages:

- Corner extraction based on automatic corner extraction from Omnidirectional Camera Calibration Toolbox for Matlab [4].
- Homography from extracted corners.
- Intrinsic and extrinsic parameters are computed from homography according to [1].
- Nonlinear optimization that minimizes the sum of the squares of the re-projection errors including the determination of distortion first for each camera separately and then for all together.

More details about calibration are described in [7].

The authors decided to make the software, including source code, publicly available. The executable, Matlab source code and sample images are available for download at http://www.ludekzalud.cz/multisenscalib/

IV. Data Fusion

Data fusion is performed by means of image transformations. The range measurements of the TOF camera can be displayed into images of the CCD cameras and thermal imagers using spatial coordinates. The thermal image can be displayed into the CCD image according to identical points (ID) of the TOF camera transformed into frames of the CCD camera and the thermal imager and vice versa (see Fig. 4).

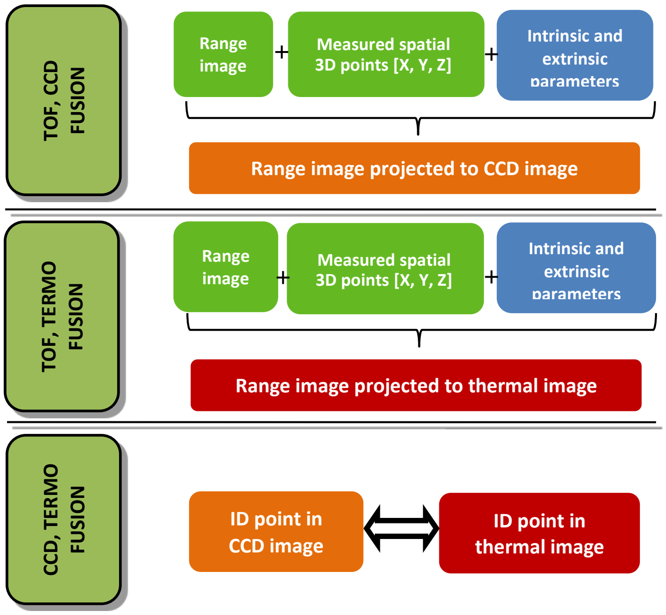

TOF, CCD FUSION
Range image
+
Measured spatial 3D points [X, Y, Z]
+
Intrinsic and extrinsic parameters
Range image projected to CCD image
TOF, TERMO FUSION
Range image
+
Measured spatial 3D points [X, Y, Z]
+
Intrinsic and extrinsic parameters
Range image projected to thermal image
CCD, TERMO FUSION
ID point in CCD image
ID point in thermal image

Fig. 4. Scheme of data fusion: TOF and CCD data fusion (up); TOF and thermal data fusion (centre); CCD and thermal data fusion (down)

The input data for data fusion include the range measurement, the image coordinates of all sensors, and the results of previous calibration. The procedure comprises the following stages:

- Computation of spatial coordinates measured by TOF camera.
- Homogeneous transformation to determine measured spatial coordinates in frames of other cameras.
- Perspective projection to determine image coordinates in frames of other cameras.
- Correction of recalculated image coordinates to the calibrated position of the principal point.

The spatial coordinates *X*, *Y*, and *Z* are computed according to (1) and (4), where *x*, *y* are image coordinates of TOF camera, *f* focal length and d_0 is measured distance projected on optical axis. Calculation of spatial coordinate *Z* in (2) is simplified by substitution of cyclometric function (3).

$$X = \frac{d_0 x}{f}, Y = \frac{d_0 y}{f}. \tag{1}$$

$$Z = d_0 = d\cos\left(\arctan\left(\frac{y}{\sqrt{f^2 + x^2}}\right)\right)\cos\left(\arctan\left(\frac{x}{f}\right)\right) \tag{2}$$

$$\cos(\tan^{-1} a) = \frac{1}{\sqrt{1+a^2}} \qquad (3)$$

$$Z = \frac{df}{\sqrt{x^2 + y^2 + f^2}} \qquad (4)$$

The homogeneous transformation is determined by (5), where $R_{[3\times3]}$ is the rotational matrix, $t_{[3\times1]}$ is the translation vector, and X', Y', Z' are the spatial coordinates of the second sensor. The image coordinates of the TOF camera in the next frame x_c',y_c' are computed using perspective projection (6), where f' is the focal length of the second sensor.

$$\begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \qquad (5)$$

$$x_c' = \frac{f'X'}{Z'}, y_c' = \frac{f'Y'}{Z'} \qquad (6)$$

A. Simulation of the most significant errors

Errors of sensors on sensor head, for that we expect the most significant impact on data fusion, were simulated. TOF camera is an indispensable element of data fusion, but less accurate than other cameras, therefore were simulated following 2 errors:

- Influence of TOF camera distance error on data fusion precision.
- Influence of low resolution of TOF camera.

The first will be discussed influence of TOF camera distance error. We determined pixel differences caused by TOF camera radial distance error for both CCD cameras and thermal imagers. We simulated distance error for 2 significant image points:

- Point on optical axis of the TOF camera.
- Point on the edge of the region 3 lying on the x-axis. Definition of TOF camera regions is determined in [8].

Measured distances in the region 4 have very low reliability, therefore this region isn´t considered. The range of the radial distance simulation is the same as detection range of TOF camera i.e. 0.1 – 10.0 m.

The effect of distance error is not significant for data fusion if transformed image coordinate differences (CCD cameras and thermal imagers) not exceed 0.5 pixel. For simulation we used values based on distance error from range calibration [7]. It is also important to judge the usefulness and impact of the range calibration. Distance error before calibration 63 mm and after calibration 30 mm was used for point on optical axis. Analogously 95 mm and 50 mm for point on the edge of reg. 3.

Fig. 5, Fig. 6, and TABLE I. show effect of pixel error in transformed images caused by distance error. Thin lines denote simulated pixel differences before TOF camera distance calibration, bold lines after. Graphs for point on the edge of reg. 3 have the same character as Fig. 5 and Fig. 6. The numerical difference is apparent from Table I.. For point on the edge of reg. 3 are pixel error slightly higher than point on optical axis (reg. 1).

Distance error is insignificant for radial distance greater than approximately 2.8 m for CCD cameras after range calibration and approximately 1.7 m in x axis and 0.3 m in y axis for thermal imagers after range calibration. Low influence of distance error on coordinate *y* of thermal imagers is caused by mounting of TOF camera and thermal imagers in the same height level on sensor head. Table I. also shows contribution of TOF camera distance calibration for more precise data fusion especially for objects in lower distances.

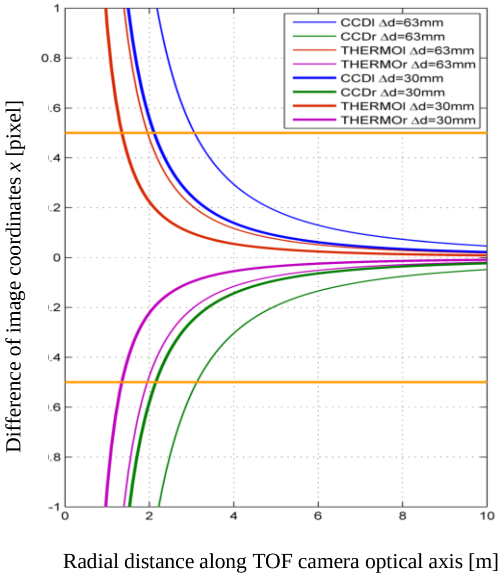


Fig. 5. Image coordinate differences Δ*x* caused by distance error for point on optical axis of TOF camera

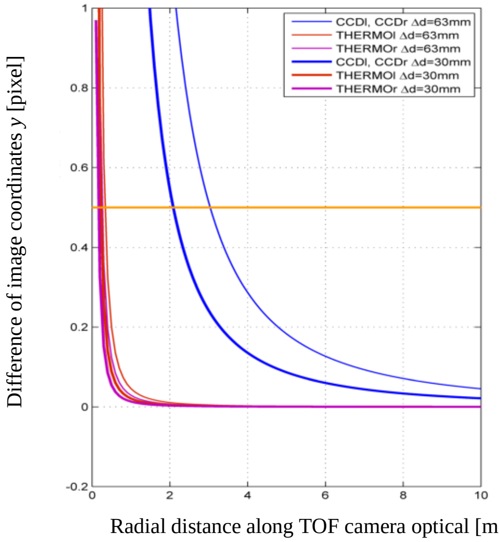


Fig. 6. Image coordinate differences Δ*y* caused by distance error for point on optical axis of TOF camera

TABLE I. IMAGE COORDINATE DIFFERENCES 0.5 PIXEL CAUSED BY TOF CAMERA DISTANCE ERROR

	Distance error	**Image coord.**	**Radial distance at that pixel error causes by distance error is 0.5 pixel [m]**			
			CCDl	***CCDr***	***TH. l***	***TH. r***
Point on optical axis	***63 mm before calibration***	*x*	3,08	3,14	1,95	1,95
		y	3,05	3,05	0,34	0,27
	30 mm after calibration	*x*	2,12	2,15	1,36	1,35
		y	2,09	2,09	0,23	0,17
Point on the edge of reg. 3	***95 mm before calibration***	*x*	3,78	3,85	2,411	2,40
		y	3,75	3,75	0,42	0,34
	50 mm before calibration	*x*	2,74	2,79	1,74	1,74
		y	2,72	2,72	0,30	0,23

Influence of low resolution of TOF camera depending on TOF camera image radial distance is the second investigated problem. Results of this simulation reflect different resolution of cameras as expected. Error 0.5 pixel in the image of TOF camera cause an error of image coordinate *x*, *y* for CCD cameras approximately 5 pixel of CCD camera and for thermal imagers approximately 2 pixel of thermal imager (see Fig. 7 and Fig. 8).

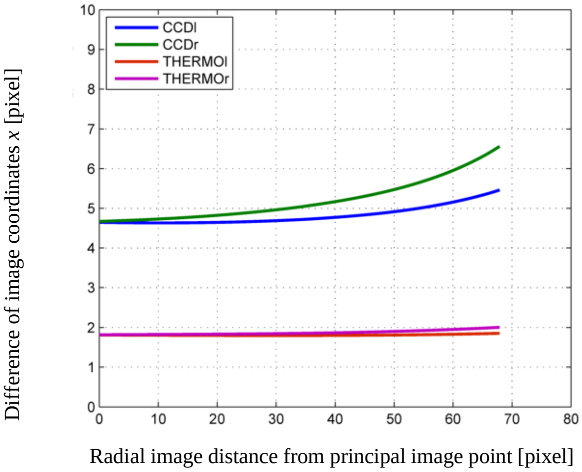


Fig. 7. Image coordinate differences Δx caused by error 0.5 pixel in TOF camera image coordinates

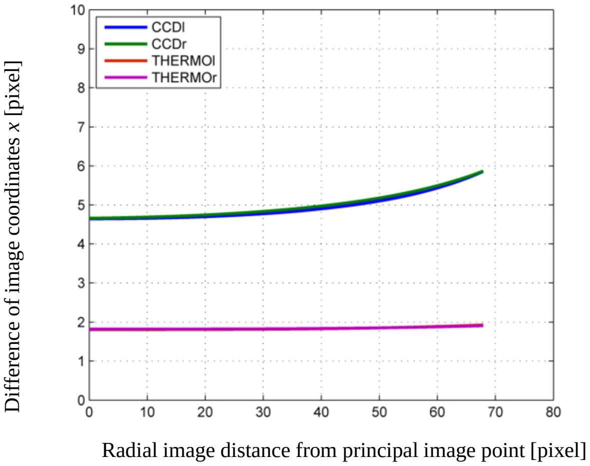


Fig. 8. Image coordinate differences Δx caused by error 0.5 pixel in TOF camera image coordinates

V. OPTIMAL IMAGE CONFIGURATION

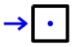
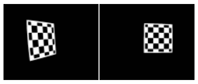
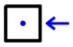
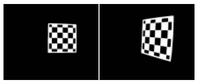
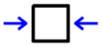
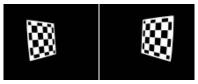
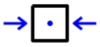
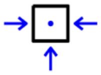
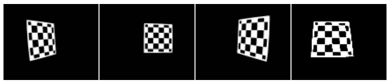
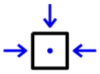
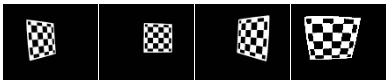
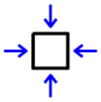
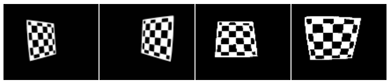
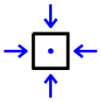
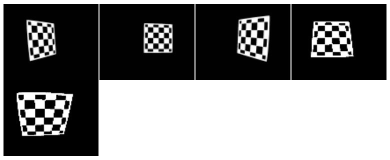
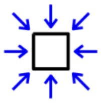
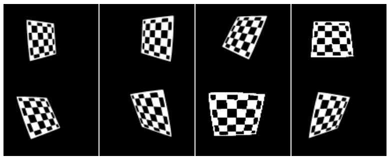
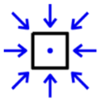
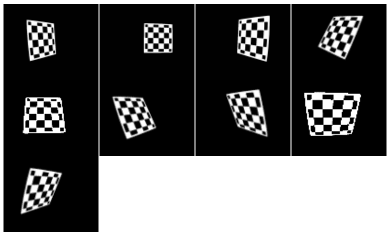
Selection of appropriate image configuration is a vital part of the whole calibration process and it has great impact to calibration results, and subsequently to multispectral data-fusion quality and robustness. To choose the most appropriate configuration, we started with 10 image configurations, see 0 Blue dots in second column denote image in normal position and blue arrows denote direction of image acquisition. 2-9 images were used for sensory head calibration. Edges of images that do not contain calibration target, are greater than usually, because of cameras rotations in sensory head and different fields of view.

The most effective configuration was determined according to independent evaluation of data fusion precision.

The principle of this evaluation is comparison of identical objects directly extracted from images from CCD cameras and thermal imagers with objects extracted from images from TOF camera and projected to CCD cameras and thermal imagers frames, using data fusion algorithm.

We had to propose objects for this verification that may be easily identifiable in the all corresponding images.

TABLE II. INVESTIGATED IMAGE CONFIGURATIONS FOR SENSORY HEAD CALIBRATION

Conf. No.	**Scheme of image acquisition**	**Examples of images for thermal imager**
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

The first our design of this object was sphere. The main reason for such choice arose from the fact that the robot moves around the objects to be identified, and it is vital that they appear identically from all points of view. The spheres can be recognized without difficulty in a color image (Fig. 9 up). In the thermal image, the identification was carried out simply via heating the metal spheres to 60°C before the measurement (Fig. 9 down). We used petanque balls (72 mm in diameter) and a shot put ball (104 mm in diameter). The most difficult problem was to recognize the spheres in the TOF camera images. Although spherical objects are commonly used for terrestrial scan registering [9], metal balls could not be reliably identified mainly due to low spatial resolution of the used TOF camera, range errors, noise, and size of the spheres.

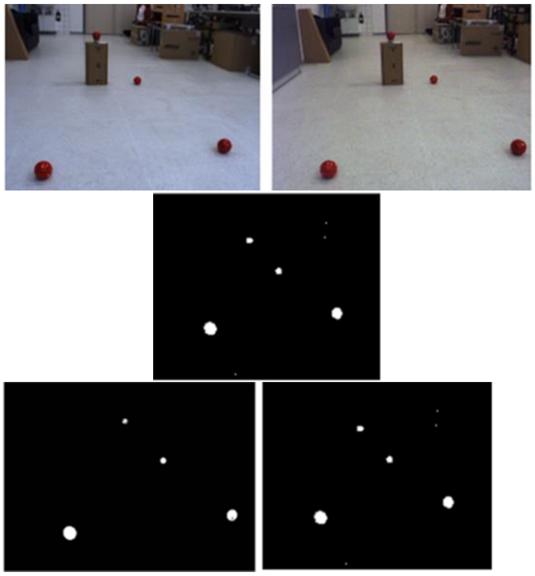


Fig. 9. The images of first target for the verification of the data fusion accuracy: the left and right CCD cameras (up); the TOF camera intensity image (center). the left and right thermal imager cameras (down)

Final design of target clearly identifiable in images of all cameras was aluminum circle covered with black paper in the middle and with 3M red reflective tape on the edge with active heating. Reflective tape is used for easier identification of targets in images of TOF camera, but significant disadvantages of this reflectivity is missing measured distances, since too big portion of light is returned unidirectional. The matte paper in the middle of the circle was used to overcome this problem – it is easy-to-be-identified by the TOF camera. We used 3 aluminum circles with 20 cm and 30 cm diameters. The targets are well identifiable on images of all 3 camera types (see Fig. 10).

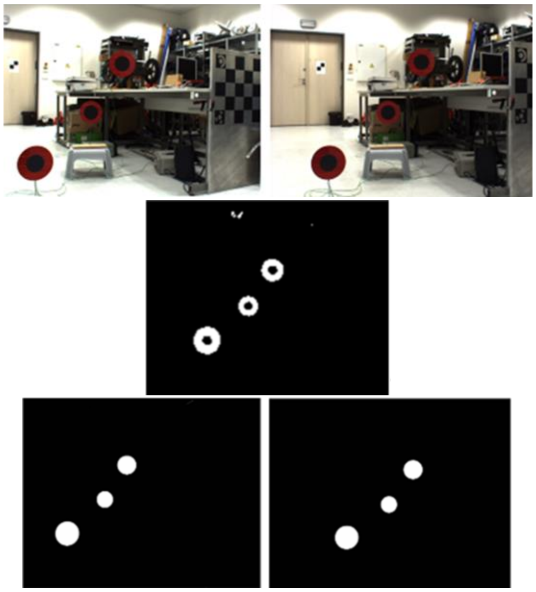


Fig. 10. The images of final target for the verification of the data fusion accuracy: the left and right CCD cameras (up); the TOF camera intensity image (center). the left and right thermal imager cameras (down)

Eighty-seven images were obtained in the experiment under real indoor conditions from the free ride of the robot. 211 extracted objects were used for data fusion evaluation, TOF camera image radial distance for these objects was in range from 1-67 pixels, range for measured distance was from 1.1 to 5.7 m.

Extraction of targets from images comprises the following stages:

- Thresholding.
- Removing small objects (noise) using morphological opening.
- Connection of separated parts using morphological closing.
- Filling closed objects.
- Determining of centroid coordinates.

0shows standard deviations σ_x, σ_y of image coordinates x, y projected by proposed data fusion algorithm for tested configurations 4-10. Standard deviation of image coordinates is denoted as σ. Values of standard deviation are given in pixels of CCD cameras and thermal imagers. Values of intrinsic and extrinsic parameters computed from only 2 images, i.e. configuration 1-3, are far away from real values.

The most suitable configurations according to values of standard deviations are configuration 8 and 10, but configuration 8 contains error increasing with image radial distance, in detail described in [7]. Configuration 10 of images is reliable according to proposed evaluation of data fusion.

TABLE III. STANDARD DEVIAATIONS OF IMAGE COORDINATES FOR CONFIGURATIONS 4-10

Conf. No		**Standard deviation od data fusion [pixel]**			
		CCDl	***CCDr***	***TH. l***	***TH. r***
4	$\boldsymbol{\sigma_x}$	6.3	6.3	2.2	3.5
	$\boldsymbol{\sigma_y}$	5.2	6.0	2.3	2.3
	$\boldsymbol{\sigma}$	5.8	6.2	2.3	3.0
5	$\boldsymbol{\sigma_x}$	2.8	3.3	1.4	1.5
	$\boldsymbol{\sigma_y}$	3.9	3.8	1.6	1.1
	$\boldsymbol{\sigma}$	3.4	3.6	1.5	1.3
6	$\boldsymbol{\sigma_x}$	2.7	3.1	1.3	1.1
	$\boldsymbol{\sigma_y}$	3.3	3.3	1.7	1.1
	$\boldsymbol{\sigma}$	3.0	3.2	1.5	1.1
7	$\boldsymbol{\sigma_x}$	3.7	3.4	1.3	1.3
	$\boldsymbol{\sigma_y}$	3.9	3.7	1.3	1.3
	$\boldsymbol{\sigma}$	3.8	3.6	1.3	1.3
8	$\boldsymbol{\sigma_x}$	2.0	2.4	1.2	1.1
	$\boldsymbol{\sigma_y}$	3.2	3.2	1.5	1
	$\boldsymbol{\sigma}$	2.7	2.8	1.4	1.1
9	$\boldsymbol{\sigma_x}$	2.9	3.5	1.2	1.1
	$\boldsymbol{\sigma_y}$	4.1	4.4	1.4	1.2
	$\boldsymbol{\sigma}$	3.6	4.0	1.3	1.2
10	$\boldsymbol{\sigma_x}$	2.4	3.0	1.0	1.2
	$\boldsymbol{\sigma_y}$	3.2	3.3	1.2	1.1
	$\boldsymbol{\sigma}$	2.8	3.2	1.1	1.2

The differences between the extracted centroid coordinates and those projected from TOF image using data fusion algorithm depending on TOF image radial distances is displayed in Fig. 11-Fig. 18for configuration 10. Due to the fact that the TOF camera has the lowest resolution, the following figure shown regions that include errors in TOF image centroid extraction in range -0.5 – +0.5 pixel (delimited by the orange horizontal lines).

The boundaries of TOF camera distance measurement accuracy regions are displayed in the following figures.

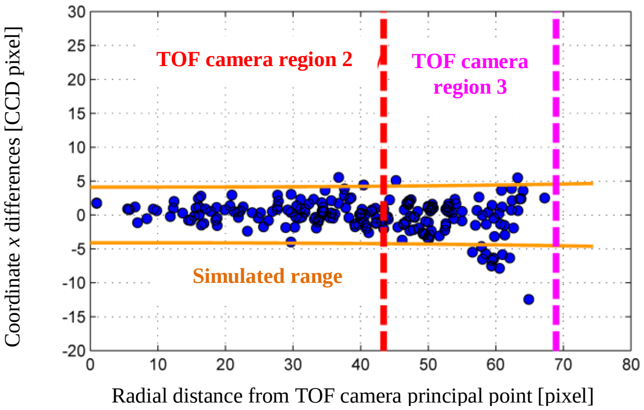


Fig. 11. The coordinate differences determined from the extracted centroids in images of the left CCD camera and from projected TOF image coordinates using the data fusion algorithm: the coordinate x differences

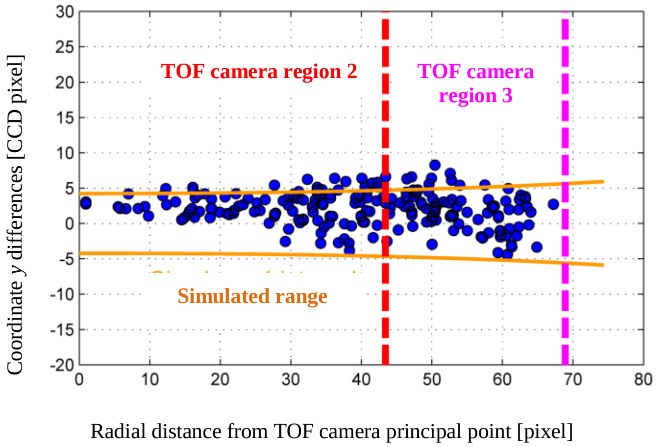


Fig. 12. The coordinate differences determined from the extracted centroids in images of the left CCD camera and from projected TOF image coordinates using the data fusion algorithm: the coordinate *y* differences

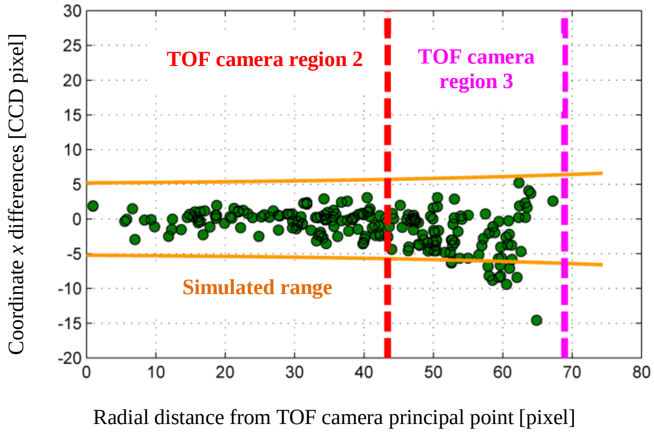


Fig. 13. The coordinate differences determined from the extracted centroids in images of the right CCD camera and from projected TOF image coordinates using the data fusion algorithm: the coordinate *x* differences

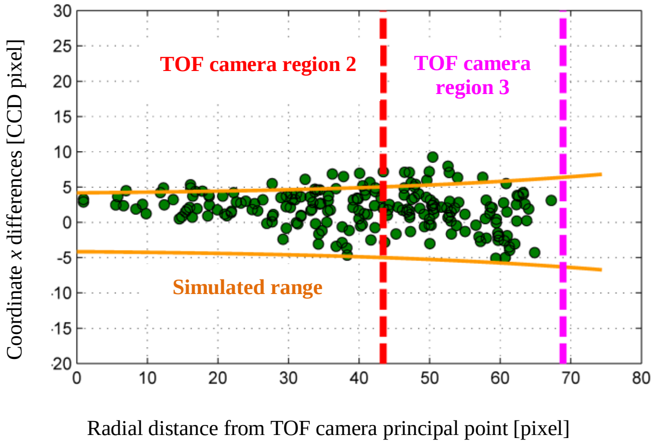


Fig. 14. The coordinate differences determined from the extracted centroids in images of the right CCD camera and from projected TOF image coordinates using the data fusion algorithm: the coordinate *y* differences

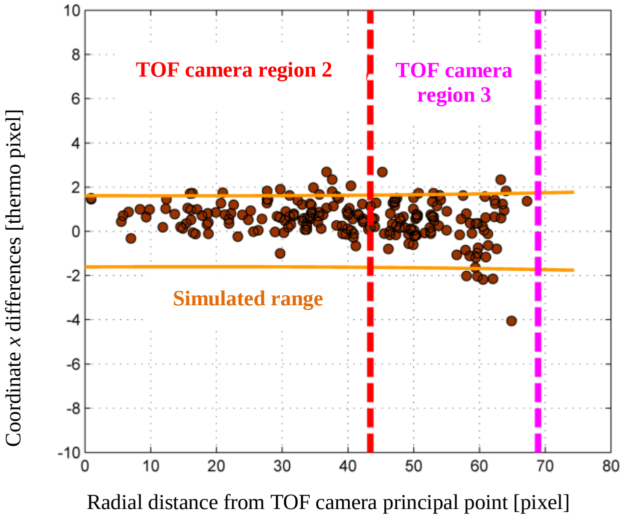


Fig. 15. The coordinate differences determined from the extracted centroids in images of the left thermal imager and from projected TOF image coordinates using the data fusion algorithm: the coordinate *x* differences

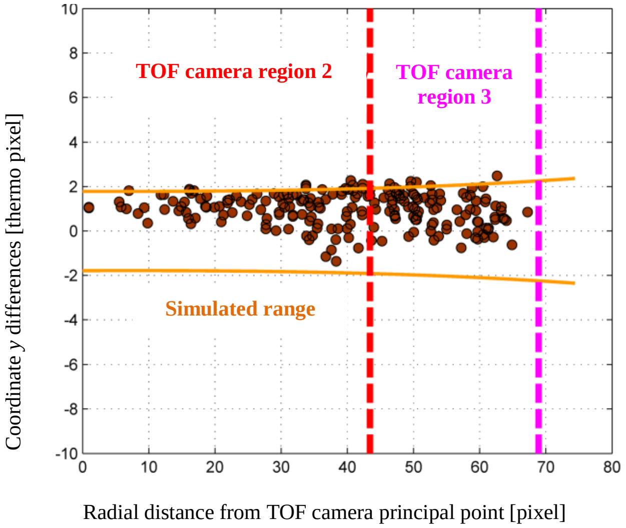


Fig. 16. The coordinate differences determined from the extracted centroids in images of the left thermal imager and from projected TOF image coordinates using the data fusion algorithm: the coordinate *y* differences

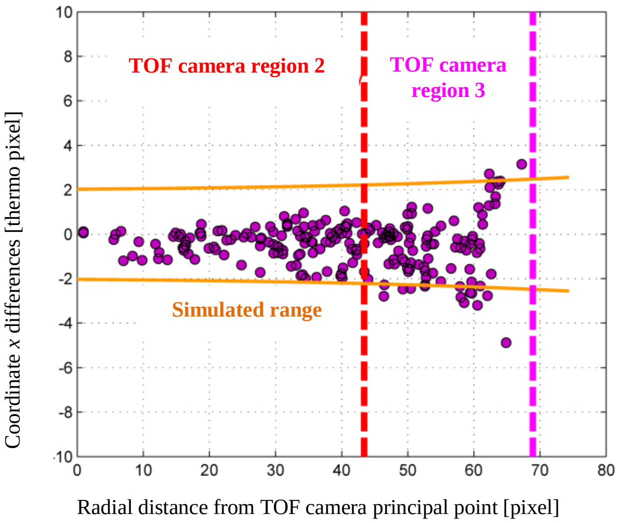


Fig. 17. The coordinate differences determined from the extracted centroids in images of the right thermal imager and from projected TOF image coordinates using the data fusion algorithm: the coordinate *x* differences

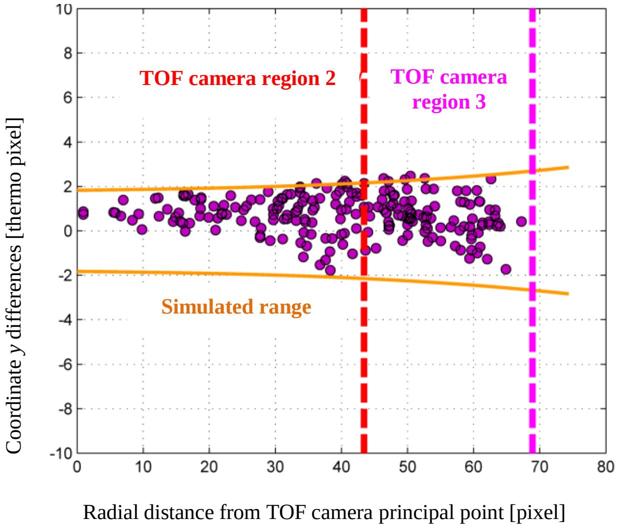


Fig. 18. The coordinate differences determined from the extracted centroids in images of the right thermal imager and from projected TOF image coordinates using the data fusion algorithm: the coordinate *y* differences

VI. Conclusion

As it is apparent from evaluation experiment described in Chapter 4, the fusion described in Chapter 3 is possible, but has its limits. The main problems come from the fact, the cameras used in the described case have significantly different spatial pixel resolution. It has to be said the cameras were carefully selected to have parameters appropriate for Orpheus-X3 robot's main mission – real-time telepresence with augmented reality containing thermal information. The cameras had to be small, lightweight, but they also offered unusually wide field-of-view. We can suppose for bigger robots sensors with considerably higher resolution might be used. The sensor resolution will also evolve in time (thermal cameras, 3D proximity cameras).

Numerical evaluation of data fusion algorithm is as follows: standard deviation for *x*, *y* image coordinates is around three pixels for CCD cameras (0.3 Pixel of TOF camera) and around 1 pixel for thermal imagers (around 0.5 TOF camera pixels).

The presented calibration process and evaluation may be used for visual and optical measurement systems of mobile robots, in general, so its use is much wider than on presented Orpheus-X3 robot demonstrator.

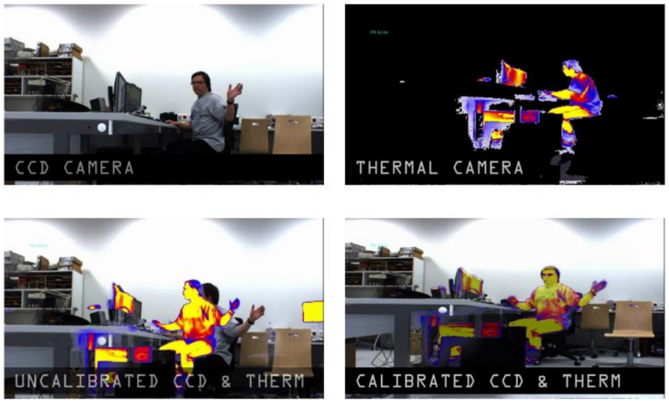


Fig. 19. Image of CCD camera (upper left), image from the thermal imager (upper right), uncalibrated data fusion (bottom left), calibrated data fusion (bottom right)

To make the calibration fast and user-friendly, we developed application MultiSensCalib in Matlab, which is available both in executable and source code in http://www.ludekzalud.cz/multisenscalib/ The same webpage also contains a set of testing images from Orpheus-X3's sensory head and a brief description of the software usage.

For the future, the authors plan to solve the biggest issue of the current data-fusion system – time latency between corresponding images. Currently if the sensory head moves rapidly, the thermal image is delayed after the camera image. Each raw data image will be equipped with time-stamp in the beginning of the data-flow and correspondingly processed. The authors also work on optimization of the algorithms to fasten the necessary processing part.

Acknowledgment

This work was supported by VG 2012 2015 096 grant named Cooperative Robotic Exploration of Dangerous Areas by Ministry of Interior, Czech Republic, program BV II/2-VS.

This work was supported by the project CEITEC - Central European Institute of Technology (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund and by the Technology Agency of the Czech Republic under the project TE01020197 "Centre for Applied Cybernetics 3".

References

[1] Z. Zhang, “Flexible camera calibration by viewing a plane from unknown orientations”, in Proceedings of the Seventh IEEE International Conference on Computer Vision. 1999. DOI: 10.1109/iccv.1999.791289.

[2] PhotoModeler Pro 5 help. Eos System Inc.: 210 - 1847 West Broadway, Vancouver BC V6J 1Y6, Canada.

[3] J.-Y. Bouguet,”Complete Camera Calibration Toolbox for Matlab” [online] http://www.vision.caltech.edu/bouguetj/calib_doc/

[4] D. Scaramuzza,”OCamCalib: Omnidirectional Camera Calibration Toolbox for Matlab” [online]. https://sites.google.com/site/scarabotix/ocamcalib-toolbox

[5] L. Zalud, F. Burian, L. Kopecny, and P. Kocmanova,” Remote Robotic Exploration of Contaminated and Dangerous Areas”, in International Conference on Military Technologies, pp 525-532, Brno, Czech Republic, ISBN 978-80-7231-917-6, 2013

[6] L. Zalud and P. Kocmanova, “Fusion of thermal imaging and CCD camera-based data for stereovision visual telepresence”, in 2013 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR). Linkoping: IEEE, 2013, pp. 1-6. DOI: 10.1109/SSRR.2013.6719344. ISBN 978-1-4799-0880-6

[7] Zalud L., Kocmanova P., Burian F. et al., “Calibration and Evaluation of Parameters in A 3D Proximity Rotating Scanner”, ELEKTRONIKA IR ELEKTROTECHNIKA, Volume 21, Issue 1, pp 3-12, DOI: 10.5755/j01.eee.21.1.7299, Kuanas univ. technol., Lithuania, ISSN 1392-1215, 2015

[8] SR4000 Data Sheet, MESA Imaging AG. Rev. 5.1, 2011.

[9] El khrachy Ismail Abd El hamid Mohamed, “Towards an automatic registration for terrestrial laser scanner data. Braunschweig”. 2007. Ph.D. thesis. Technische Universität Carolo-Wilhelmina.

[10] Zalud L., Burian F., Kopecny L., Kocmanova P., “Remote Robotic Exploration of Contaminated and Dangerous Areas“, International Conference on Military Technologies, pp 525-532, Brno, Czech Republic, ISBN 978-80-7231-917-6. (2013)

[11] Zalud L., Kopecny L., Burian F., “Robotic Systems for Special Reconnaissance“, ICMT'09: INTERNATIONAL CONFERENCE ON MILITARY TECHNOLOGIES, pp 531-540, Brno, Czech Republic, ISBN 978-80-7231-649-6, 2009

[12] Nejdl L., Kudr J., Cihlarova K. et al, “Remote-controlled robotic platform ORPHEUS as a new tool for detection of bacteria in the environment”, Electrophorensis, Volume 35, Issue 16, pp 2333-2345, Wilwy-Blackwell, USA, DOI 10.1002/elps.201300576, ISSN 0173-0835, 2014

Optimizing User's Utility from Cloud Computing Services in a Networked Environment

Eli WEINTRAUB
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

Yuval COHEN
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

***Abstract*—Cloud Computing customers are looking for the best utility for their money. Research shows that functional aspects are considered more important than service prices in customer buying decisions. Choosing the best service provider might be complicated since each provider may sell three kinds of services organized in three layers: SaaS (Software as a service), PaaS (Platform as a service) and IaaS (Infrastructure as a service). This research targets the problem of optimizing consumers' utility, using conjoint analysis methodology. Providers currently offer software services as bundles belonging to the same layer, or to underlying layers. Bundling services prevent customers from splitting their service purchases between a provider of software and a different provider of the underlying layers. This research assumes that in the future will exist a free competitive market, in which consumers will be free to switch their services to different providers, eliminating the negative biases of bundling, during making their buying decisions. This research proposes a mathematical model and three possible strategies for implementation in organizations, and illustrates its advantages compared to existing utility maximization practices. Current conjoint analysis method chooses the best utility in a traditional cloud architecture in which one provider offers a bundle of all three layers. The proposed model assumes a networked cloud architecture in which a customer may choose services from any provider, building for himself the best basket of services maximizing his/her total utility. This research outlines three business models which will assist organizations shift gradually from current CC architecture to the future networked architectures, thus maximizing their utility.**

***Keywords*—Utility Optimization; Cloud Computing; Consumer preferences; Conjoint Analysis**

I. Introduction

In the last years organizations began to shift parts of their computing infrastructures outside the geographic organizational borders to the cloud, to other organization which owns the infrastructure. Ref. [15] states that shifting computing facilities outside the organizations' borders enforces establishing new processes of production control, service level monitoring implementing solutions to security and privacy issues. Most definitions of Cloud Computing (CC) state that it’s a technology enabling on-demand services, scalability, and flexibility, in computing consumption ([19] [18]). The National Institute of Standards and Technology (NIST) defines CC as a model which enables convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly released with minimal service provider interaction [13]. Ref. [12] argues that from time to time cloud providers suffer outages, thus contacting a multi-cloud broker is a preferred solution to keep high up time of services. Service brokers also diminish interfacing efforts needed to various protocols used by service providers. Ref. [1] suggest adding an Inter-Cloud computing layer to CC systems, which enables to shift resources among cloud systems, thus improving downtime measures and computing resource management. A. Gill, D. Banker, and P. Seltsika [5] who studied the future potential financial services technologies found that CC is a cost-effective infrastructure compared to traditional infrastructures.

According to [22] CC organizational adoption models can be classified into four types: private, public, community and hybrid. Organizations which adopt the private model, locate its infrastructures outside the organization’s sites under the responsibility of a cloud service provider. In a public model, the organization chooses a cloud service provider having the best proposal among cloud public service providers. A public cloud computing provider usually uses the same computing infrastructures for other companies. In a community adoption model, infrastructure services are shared by a group of customers. In a hybrid model, organizations can use infrastructure services supplied by public, private or community providers.

Consumers who wish to use CC services have to decide the selection criteria for evaluating service providers' services. Such a selection might be complicated to measure and compare since providers offer different services having various functionalities, on un-standard scales. Conjoint analysis has been cited in literature as a methodology which enables coping with providers' selection issue. Conjoint analysis has been demonstrated on a common CC architecture which limits consumers' choices. This research suggests using the conjoint analysis methodology implemented on a networked CC architecture. The research presents a mathematical model and three business strategies which enable maximizing consumers' utility compared to existing business model. The suggested CC architecture is aimed at a future free market competition, in which consumers will be free to choose their service providers improving their utility measures.

The rest of the paper is organized as follows: Section II describes the conjoint analysis methodology, and describes a survey performed, which studied cloud computing consumers' preference attributes. This research defines an optimization model, making use of findings from that previous survey. Section III describes a cloud computing architecture according to the current common usage, and according to the new dynamic networked model suggested in literature. Section IV presents and demonstrates the optimization model in the current architecture, and also presents three suggested optimization models implemented on the dynamic architecture. Section V presents and discusses research findings, and a compares the three suggested optimization models. Finally, section VI suggests future research directions.

II. Conjoint Analysis in Cloud Computing

Conjoint analysis is a methodology often used by researchers, aimed at analyzing buyers trade-offs among competing products [6]. The analysis enables simulating and predicting buyers' considerations when they compare different products looking at the characteristics each product resembles. Ref. [2] used the methodology by studying consumers' buying consideration of CC services, and found that most influencing CC service characteristics are service availability advantages and lock-in prevention. The authors found that consumers did not name cost savings as a major factor influencing their buying decision. P. Koehler, A. Anandasivam, and A. Dan [9] who analyzed consumers' decisions using CC services found that consumers have not mentioned cost savings as a major consideration. According to Ref. [4], information security has been found a barrier to CC adoption. According to Ref. [17], information security is today a barrier to CC adoption but in the future security will not be a barrier, since information security technologies will become less technological and more effective. Ref. [20] foresees a shift from technology issues to an emphasis on service-based consideration in customer value using CC services.

In the survey performed by P. Koehler, A. Anandasivam, M. Dan, and C. Weinhardt, [8] the researchers asked consumers' about service attribute preferences, and found that the consumers named six attribute levels. The researchers then performed a choice based conjoint analysis and concluded that the attributes were: (1) providers' reputation, (2) required skills, (3) migration process, (4) pricing tariff, (5) cost compared to internal solution and (6) consumer support. Consumers have not mentioned security as a preference attribute. Although security is considered a barrier to adopting CC services as stated above, Ref. [21] states that consumers are un-capable of evaluating the differences in cloud security services. Ref. [8] used conjoint analysis by comparing the relative importance of customers' decision attributes, and found out that provider reputation was the attribute with the highest relative importance of 26% out of all other attributes, and migration process was the second most important attribute with 21% importance. The cost has been found only in the fourth place having 16% relative importance. Table II lists six attribute preference importance levels computed in the research using conjoint analysis methodology**.** W. Venters and E.A. Whitley [20] who studied the attributes influencing on customers decisions, claim that consumers do not consider CC as an alternative delivery and pricing mechanism, but as a tool that enables creative use of technology for achieving business targets. In Ref. [3] researchers studied the service attributes influencing on CC adoption. They found seven groups of attributes: Monetary payoff, usability, flexibility, trademark, added value, connectivity and customers' support.

The paper suggests a new model that maximizes consumers' utility in a multi-services providers' environment. The proposed model simulates consumers' choice behavior by finding the maximal utility, assuming that the consumer evaluates his utility by using conjoint analysis technique. This paper develops a model that will enable understanding customers' buying decisions in cloud computing services, assuming that each SP's attributes are given, as previous consumers pointed out during visiting SP's websites. The research also assumes that the consumer can select all combinations of cloud services (SaaS, PaaS, and IaaS) from different providers.

III. Cloud Computing Architectue

Research literature describes cloud computing architecture as consisting of three layers: IaaS, PaaS and SaaS. Each layer performs certain functions, serving consumers' requests. This separation to layers also fits current services offered by cloud providers. Ref. [22] defines a framework of CC architecture composing three layers of functions supporting cloud computing services. Systems' architectures' components are outlined in Fig. I. White rectangles describe computing services, grey rectangles describe computing resources. Following the functions performed by each layer.

Infrastructure layer – focuses on providing technologies as basic hardware components for software services. There are two kinds of infrastructures: storage capabilities and computing power. Platform layer - includes services which are using cloud infrastructures needed for their functioning. There are two kinds of platform services: development and business platforms. Development platforms are aimed for usage by developers who write programs before transferring them to production and usage by organizations' users. Business platforms enable organizational developers make adaptations of software packages for deployment in their organizations. Application layer - consists of programs and human interfaces used by the organizations' end-users. Applications are running on cloud assets, making use of platform and infrastructure layers. There are two kinds of services in this layer: applications and on-demand services. Application services are software packages ready for end-users such as Microsoft Office, while on-demand services are software applications used by the organizations' customers. Those services are used according to on-demand needs, and used on a pay-per-use or fixed-price pricing model.

To summarize, Service Providers (SP) offer their customers' three kinds of services: IaaS, PaaS and SaaS. Each SP manages all underlying infrastructure for the offered service. For example a SP suggesting a SaaS product usually bundles into the product the PaaS and IaaS layers. Ref. [19] states that according to cloud computing architecture a certain provider may run an application using another provider’s infrastructure, but in practice both providers are parts of the same organization. According to current practice, when a provider suggests selling a PaaS service he also bundles the IaaS layer in the deal. Such bundling by service providers limit free market forces from entering the competition, forcing customers pay for components they may buy cheaper from other providers. For example a customer may buy a SaaS service from SP1, but buy the underlying PaaS service from SP2 which sells the appropriate platform service cheaper than SP1. Ref. [14] claims that in the future, developers will plan their cloud applications enabling migration of services among clouds of multiple clouds. According to Ref. [24] cloud computing architecture is more modular compared to traditional hosting architectures. CC is based on server farms, with programs running on different layers which are loosely coupled, thus enabling the development of a wide range of applications. Ref. [19] claims it is possible that applications belonging to different layers will run on separate geographical locations even in different countries. Ref. [16] claims that virtual machine migration allows transfer of a running application from one virtual machine to another, which may be provided by a different IaaS provider.

This Research assumes existence of a business model which enables implementing the needed functionalities of a service provider which operates the underlying platform using other service providers, according to consumers' preferences. Implementing this required functionality puts two requirements on cloud architecture. Firstly, the architecture should be based on open standards which will enable interfacing between many components among providers offering all three layers. Second, the architectures' building blocks should be loosely coupled. Implementation of those two functionalities should enable connectivity among vertical and horizontal services, thus elimination of the bundling phenomena defined by E. Weintraub and Y. Cohen [23]. They introduced new definitions of two kinds of bundling: first is horizontal bundling, second is vertical bundling. In horizontal bundling a provider offers several services, all belong to one layer. For example Amazon EC2 offers several bundles each one is composed of the following components: CPU, ECU, memory, instance storage, and operating system. In such bundling situations consumers may not use their own operating system. In vertical bundling a provider offers services which belong to lower layers, in addition to the main needed service. For example Amazon offers SaaS services, in which the consumer is asked to choose the configuration of infrastructure he wants the software application to run. A consumer may not use a PaaS service such as his own operating system. Figure II describes the suggested dynamic cloud architecture. Arrows describe services supplied by underlying layers. Rectangles describe computing services.

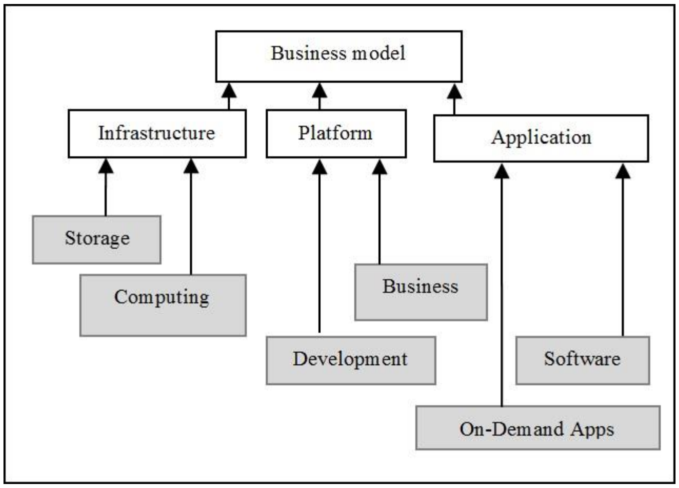


Fig. 1. Current Cloud business model Architecture

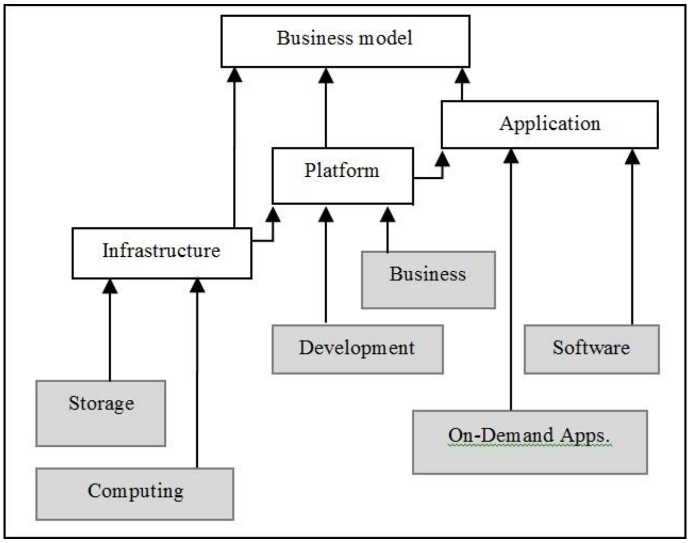


Fig. 2. A Dynamic Architecture for Cloud Computing from Wentraub and Cohen (2015)

IV. CONSUMERS' UTILITY OPTIMIZATION ACCORDING TO CURRENT MODEL

In an ideal world the optimal service selection is a combination of the maximal utility of each attribute.

According to current practices most combinations are not feasible or not offered since CC providers offer bundles of services, raising difficulties on consumers wishing to buy certain services from another provider, thus limiting their dependence on the main CC provider. In order to estimate consumers' preferences, a choice based conjoint analysis, which was first introduced by J.J. Louviere and G.G. Woodworth [11] was included in the survey described in Ref. [9], analyzing service attributes and attribute levels for describing cloud services. Following R. Weiber and d. Mühlhaus [21], A. Hollobaugh [7] generated a list of 18 attributes with 49 attributes' levels. A. Hollobaugh reduced this list by validating it through expert interviews resulting in the final selection of 6 attributes as detailed in Table I [8].

TABLE I. CONSUMERS' SELECTION ATTRIBUTES FROM [8]

Selection Attribute	Explanation
Provider Reputation	The reputation of the service providers refers to the attitude, beliefs and trust.
Required Skills	Do consumers need to be trained or have specific skills in order to use the services, can easily use services.
Migration Process	Can users use standard data formats or they have to use provider specific data formats.
Pricing Tariff	Are services offered with a pay-as-you-use tariff, or with a flat rate tariff in which they can use the service as often as they want or a one-time-purchase in which the consumer only pays an initial price at the first time of use and can use it unlimited.
Cost compared to internal solution	In that survey, Cloud services can have equal costs compared to an intern solution, but also may have 15% less or 25% less costs.
Consumer support	Do providers offer consumer support in different ways such as FAQ, email, forums etc' if their consumers need help.

P. Koehler, A. Anandasivam, M. Dan, and C. Weinhardt [8] created efficient choice sets using SAS, based on W. F. Kuhfeld [10] model. In total, 13 choice sets have been included in the survey, and each consists of three alternatives and one non-choice option. The researchers estimated the part worth utilities for all attributes, except cost reduction, based on Ref. [10] model. The choice decision behavior was predicted using a multinomial Logit choice model (MNL). Afterwards, the part worth utilities were estimated with a maximum likelihood approach (ML) and finally were standardized to relative importance. The null hypothesis that "there are no strictly preferred attributes and all part worth utilities equal zero" was rejected based on a significance level of α=0.01 and a likelihood ratio of 161.34. The researchers found that potential consumers of cloud services do have a strong preference with regards to the different service attributes.

TABLE II. CONSUMERS' ATTRIBUTE IMPORTANCE PREFERENCES FROM [8]

Selection Attribute	Importance %
Provider reputation	26
Required skills	7
Migration process	21
Pricing tariff	17
Cost compared to internal solution	16
Consumer support	13

Table II shows the relative importance of all attributes' levels (out of the total sample). Suppose an ideal computing service gives a consumer the maximum utility (100%). Such a cloud service should be offered by a provider with high reputation (26% relative importance, which is the highest one), operates a smooth migration process (21%), and requires no additional training. From an economic point of view, respondents prefer a flat rate pricing model (17%) and as much cost reduction as possible (16%). Corresponding IT support is offered preferably by standard electronic sources (13%), such as frequently asked questions and documentation. The results provide an idea how consumers' utility reacts to differing some service attributes.

Following, this paper describes a theoretical use case of three SP's and their tariffs, and an organizational customer who wishes to choose certain SP's, based on his business and computing requirements. This paper presents three pricing models which make use of the data described in both previous section and this section.

A. Provider's bundle

In this section, following a definition of a mathematical model which helps the users in their decision finding the vendor proposing the maximal utility. The model makes use of CC service attributes that contribute value to customers as suggested by [8]. The proposed model is based on the attributes and weights presented in section IV (summarized in tables I and II). The comparison shall use the following use case. A consumer with a list of needed CC services is considering three candidate service providers (named SP1, SP2, SP3) in order to choose a service provider (for this list of services). The consumer must acquire some information related to service qualities of each provider. Some sources of data may be (1) consulting companies, (2) forums, providers (3) white papers and (4) customers' experience related to specific SP. Then, the consumer has to rank the relative grade of the service attributes of each provider. This is, for each service such as "Data transfer" or "Email services" and for each SP, the customers rank the five attributes: (1) Provider reputation, (2) Required skills (3) Migration process (4) Pricing tariff (5) Cost compared to internal solution (6) Consumer support.

The use case defines a scale of numbers from 1 to 3 (where 1 the worst utility level, and 3 is the highest utility level). Of course any scale is acceptable and some would feel more comfortable with a Likert scale of 1 to 5. For example, Data transfer service for SP1 is graded 132213 yielding an average grade of 59% detailed in Table III.

Table IV details the grades for this case study for three SP's. The meaning of the upper left grading (132213) is detailed in Table III.

As a demonstration of the model, the use case assumes a user who wants to optimize his utility for the following attribute selection and service levels detailed in Table VI.

The users' general objective is to choose the maximal utility solution for the list of required services. The underlying assumption is that the utilities of all services are computed in the same way by the consumer.

TABLE III. AN EXAMPLE OF GRADING A SERVICE PROVIDERS' (SP1) SERVICE

Selection Attribute	Grade	Importance %	Weighted AVG
Provider reputation	1	26	26
Required skills	3	7	21
Migration process	2	21	42
Pricing tariff	2	17	34
Cost compared to internal solution	1	16	16
Consumer support	3	13	39
Total			**178**
Final grade in %			**178/300=59%**

This sub-section is devoted to the first model out of three different maximal utility models to be described in the next sub-sections. The models are based on three architecture configurations according to which the optimal solution is chosen. The first model focuses on choosing a single supplier out of the three in the case study.

TABLE IV. SERVICE UTILITIES OF THREE PROVIDERS. EACH UTILITY IS RANKED BY A GRADE 1,2 OR 3. SIX SERVICE VALUES OF EACH PROVIDER ARE RANKED USING A VECTOR CONTAINING 6 NUMERIC UTILITY GRADES

SaaS	Service name	SP1 Utilities	Sp2 Utilities	SP3 Utilities
	Data transfer	132213	133321	232323
	Email services	332211	221213	333323
	Cloud search	332132	211132	123331
	Documents Mgt.	131321	231231	312321
	ERP	332313	111212	332131
PaaS	Service name	SP1 Utilities	Sp2 Utilities	SP3 Utilities
	Operating system	112211	323213	311121
	Memory	122132	233132	121131
	Instance storage	131121	231231	312321
	Developer support	332323	311212	332131
IaaS	Service name	SP1 Utilities	Sp2 Utilities	SP3 Utilities
	Relational Database services	112213	133321	232323
	Storage standard vol.	122211	323213	311123
	Backup	112132	211132	123331

According to the Providers bundle model the consumer has to choose one SaaS provider which adds to his bundle PaaS and IaaS services. This model is characterized by both horizontal and vertical bundling. The consumer has no possibility to choose any service other than the Chosen SP. The calculations start by computing the total utility of each service of each SP. Maximizing consumers' utility in this model is implemented by choosing the SP with the highest utility summing the service utility of all three kinds of services: SaaS, PaaS and IaaS.

Total Utility for the various SPs (see computations below) is: (SP1) = 22.04, SP2 = 23.35, SP3=25.73

Thus, SP3 (with highest utility) is chosen. The following is the detailed computation:

Total Utility (SP1) =

0.26x (1+3+3+1+3+1+1+1+3+1+1+1) +
0.07x (3+3+3+3+3+1+2+3+3+1+2+1) +
0.21x (2+2+2+1+2+2+2+1+2+2+2+2) +
0.17x (2+2+1+3+3+2+1+1+3+2+2+1) +
0.16x (1+1+3+2+1+1+3+2+2+1+1+3) +
0.13x (3+1+2+1+3+1+2+1+3+3+1+2) =
=**22.04**

Total Utility (SP2) =

0.26 x (1+2+2+2+1+3+2+2+3+1+3+2) +
0.07x (3+2+1+3+1+2+3+3+1+3+2+1) +
0.21x (3+1+1+1+1+3+3+1+1+3+3+1) +
0.17x (3+2+1+2+2+2+1+2+2+3+2+1) +
0.16x (2+1+3+3+1+1+3+3+1+2+1+3) +
0.13x (1+3+2+1+2+3+2+1+2+1+3+2) =
= **23.35**

Total Utility (SP3) =

0.26 x (2+3+1+3+3+3+1+3+3+2+3+1) +
0.07x (3+3+2+1+3+1+2+1+3+3 +1+2) +
0.21x (2+3+3+2+2+1+1+2+2+2+1+ 3) +
0.17x (3+3+3+3+1+1+1+3+1+3+1+ 3) +
0.16x (2+2+3+2+3+2+3+2+3 +2+2+3) +
0.13x (3+3+1+1+1+1+1+1+1 +3+3+1) =
= **25.73**

To conclude, SP3 is chosen as best utility supplier for the consumer, producing **25.73** utility units.

Following, the paper presents a demonstration of choosing the best solution of a CC Hierarchical pricing model.

B. *Hierarchical model*

The Hierarchical model assumes the consumer may choose different service providers, but limiting each SP to supply all services requested in each layer. Thus, the vertical bundling constraint is released, but the vertical bundling constraint is still valid. Since fitting SaaS services to the customer is more sensitive to customer requirements (and usually more expensive) – this model assumes that each SP maximizes its SaaS capabilities and looks for purchasing the best combination of platform and infrastructure services that best complements its own offerings in these levels. Since SPs seek simple management and control of sub-contracted services, only one SP could be chosen for complementing the platform or the infrastructure level. The Platform SPs can also purchase infrastructure services. While SaaS is the highest level in the hierarchy, the computations start from the lowest level (IaaS) and progress through PaaS to the decision taken by the SP based on their SaaS and possibly sub-contracted PaaS and/or IaaS. The customer in this model would choose at each level the provider of choice for the requirements.

1) *Fist Hierarchical Level*

The calculations start with comparing the infrastructure services of the three candidate SPs, as follows:

IaaS Utility (SP1) =

0.26 x (1+1+1) +
0.07x (1+2+1) +
0.21x (2+2+2) +
0.17x (2+2+1) +
0.16x (1+1+3) +
0.13x (3+1+2) =
=**4.75**

IaaS Utility (SP2) =

0.26 x (1+3+2) +
0.07x (3+2+1) +
0.21x (3+3+1) +
0.17x (3+2+1) +
0.16x (2+1+3) +
0.13x (1+3+2) =
= **6.21**

IaaS Utility (SP3) =

0.26 x (2+3+1) +
0.07x (3+1+2) +
0.21x (2+1+3) +
0.17x (3+1+3) +
0.16x (2+2+3) +
0.13x (3+3+1) =
= **6.46**

Hence SP3 is chosen as best IaaS utility supplier for the consumer, producing **6.46** utility units.

Now PaaS has to be evaluated for the three SPs.

2) Second Hierarchical level

The calculations start with comparing the platform services of the three candidate SPs, as follows:

PaaS Utility (SP1) =

0.26 x (1+1+1+3) +
0.07x (1+2+3+3) +
0.21x (2+2+1+2) +
0.17x (2+1+1+3) +
0.16x (1+3+2+2) +
0.13x (1+2+1+3) =
=**7.04**

PaaS Utility (SP2) =

0.26 x (3+2+2+3) +
0.07x (2+3+3+1) +
0.21x (3+3+1+1) +
0.17x (2+1+2+2) +
0.16x (1+3+3+1) +
0.13x (3+2+1+2) =
= **8.42**

PaaS Utility (SP3) =

0.26 x (3+1+3+3) +
0.07x (1+2+1+3) +
0.21x (1+1+2+2) +
0.17x (1+1+3+1) +
0.16x (2+3+2+3) +
0.13x (1+1+1+1) =
= **7.49**

If platform level SP is chosen independent of other levels SP2 would be chosen as best utility supplier for our consumer, producing **8.42** utility units.

This model assumes now the need of an interfacing fee for connecting a platform held by service provider *i* to an infrastructure *j* when the infrastructure belongs to a different service provider. As an example the model assumes that the interfacing fee is worth **0.05** utility units. The model computes now all combinations of a platform service provider and an infrastructure service provider producing maximum utility.

Platform SP1 + Infrastructure SP1 = 7.04 + 4.75 = 11.79

Platform SP2 + Infrastructure SP2 = 8.42 + 6.21 = 14.63

Platform SP3 + Infrastructure SP3 = 7.49 + 6.46 = 13.95

Platform SP1 + Infrastructure SP2 – interface fee = 7.04 +6.21 – 0.05 = 13.2

Platform SP1 + Infrastructure SP3 – interface fee = 7.04 + 6.46 – 0.05 = 13.45

Platform SP2 + Infrastructure SP1 – interface fee = 8.42+ 4.75 – 0.05 = 13.12

Platform SP2 + Infrastructure SP3 – interface fee = 8.42+ 6.46 – 0.05 = 14.83

Platform SP3 + Infrastructure SP1 – interface fee = 7.49+ 4.75 – 0.05 = 12.19

Platform SP3 + Infrastructure SP2 – interface fee = 7.49+ 6.21 – 0.05 = 13.65

Hence Maximum utility is achieved combining Platform SP2 with IaaS SP3 producing **14.83** utility units.

As the interface cost grows, the solutions without interfaces are more attractive.

For example, an interface fee of 0.3 would yield:

Platform SP2 + Infrastructure SP3 – interface fee = 8.42+ 6.46 – 0.3 = 14.58

whereas **Platform SP2 + Infrastructure SP2 = 8.42 + 6.21 = 14.63**

So SP2 without any interface becomes the chosen alternative.

Another example assuming no interface fee using the hierarchical model calculating utility:

Platform SP1 + Infrastructure SP1 = 7.04 + 4.75 = 11.79

Platform SP2 + Infrastructure SP2 = 8.42 + 6.21 = 14.63

Platform SP3 + Infrastructure SP3 = 7.49 + 6.46 = 13.95

Platform SP1 + Infrastructure SP2 = 7.04 + 6.21 = 13.25

Platform SP1 + Infrastructure SP3 = 7.04 + 6.46 = 13.50

Platform SP2 + Infrastructure SP1 = 8.42 + 4.75 = 13.17

Platform SP2 + Infrastructure SP3 = 8.42 + 6.46 = 14.88

Platform SP3 + Infrastructure SP1 = 7.49 + 4.75 = 12.24

Platform SP3 + Infrastructure SP2 = 7.49 + 6.21 = 13.70

Hence Maximum utility is achieved combining Platform **SP2** with IaaS SP3 producing **14.88** utility units.

Now the model considers SaaS and selects the best software service provider.

3) Third Hierarchical Level

SaaS Utility (SP1) =

0.26 x (1+3+3+1+3) +
0.07x (3+3+3+3+3) +
0.21x (2+2+2+1+2) +
0.17x (2+2+1+3+3) +
0.16x (1+1+3+2+1) +
0.13x (3+1+2+1+3) =
= **10.25**

SaaS Utility (SP2) =

0.26 x (1+2+2+2+1) +
0.07x (3+2+1+3+1) +
0.21x (3+1+1+1+1) +
0.17x (3+2+1+2+2) +
0.16x (2+1+3+3+1) +
0.13x (1+3+2+1+2) =
= **8.72**

SaaS Utility (SP3) =

0.26 x (2+3+1+3+3) +
0.07x (3+3+2+1+3) +
0.21x (2+3+3+2+2) +
0.17x (3+3+3+3+1) +
0.10x (2+2+3+2+3) +
0.11x (3+3+1+1+1) =
= 11**.78**

From previous stages the model choses the maximal SaaS provider incorporating Platform **SP2** and Infrastructure **SP3** whose utility is **14.83** as computed.

Software SP1 + 14.83 – 0.05 = 10.25 +14.83 - 0.05 = 25.03

Software SP2 + 14.83 = 8.72 + 14.83 = 23.55

Software SP3 + 14.83 – 0.05 = 11.78 +14.83 - 0.05 = 26.56

Hence the model chooses as maximal utility architecture producing **26.56** utility units, offering a combination of service providers: **Software SP3, Platform SP2 and Infrastructure SP3**.

This selection produces an improved utility **26.56** over the providers' bundle model which chose SP3 as maximal utility architecture producing only **25.73** utility units. The improved utility is achieved in the hierarchical model in spite of the payoff of two interfacing fees paid for connection software SP3 to platform SP2, and secondly connecting platform SP2 to infrastructure SP3. In this example the model demonstrated using the hierarchical model achieving higher overall utility by using the flexibility of choosing layers of service providers free of hierarchical bundling constraints.

Following, the use case assumes that no interface fee is needed for calculation of the best alternative:

Software SP1 + 14.88 = 10.25 +14.88 = 25.13

Software SP2 + 14.88 = 8.72 + 14.88 = 23.60

Software SP3 + 14.88 = 11.78 +14.88 = 26.66

Hence, the model chooses as maximal utility architecture producing **26.66** utility units, offering a combination of service providers: Software SP3, Platform SP2 and Infrastructure SP3.

Following, the model presents the third architecture model demonstrating achieving even higher utility by relaxing the horizontal bundling constraints.

C. Optimized model

In this model the consumer may choose services freely in a free competitive market, selecting the best service in the market according to the utility gained from the service. In this model both vertical and horizontal bundling are relaxed, and the consumer may choose each service under no constraints whatsoever. The model introduces three sub-models according to fees management strategies.

In this section the paper presents an analysis of the impact of the cost of administrative work (ordering, tracking and payment management) on the optimal policy in a free market setting. First, the paper introduces the simple basic model without fees or costs. Then, the paper presents a maximal utility approach. Finally the paper presents the direct utility comparison for choosing a primary supplier.

1) The Basic Optimized Model

In this utility model the research assumes free market rules, in which each service is chosen to be the one that brings the highest utility. Table IV summarizes this utility model:

TABLE V. BASIC OPTIMIZED UTILITY MODEL

SaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	Max Utility	Best SP
	Data transfer	1.78	2.06	2.37	2.37	3
	Email services	2.04	1.76	2.84	2.84	3
	Cloud search	2.32	1.71	2.15	2.32	1
	Documents Mgt.	1.64	1.89	2.23	2.23	3
	ERP	2.47	1.30	2.19	2.47	1
PaaS	Service name	SP1 Utility	Sp2 Utility		Max Utility	Best SP
	Operating system	1.38	2.44	1.68	2.44	2
	Memory	1.73	2.27	1.39	2.27	2
	Instance storage	1.30	1.89	2.23	2.23	3
	Developer support	2.63	1.82	2.19	2.63	1
IaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	Max Utility	Best SP
	Relational Database services	1.64	2.06	2.37	2.37	3
	Storage standard vol.	1.45	2.44	1.94	2.44	3
	Backup	1.66	1.71	2.15	2.15	3
Total					**28.76**	

Ignoring the cost of managing multiple SPs, the total utility in this case would be the sum of the Max utility column: **28.76**. This is of course a better utility than the other two utility methods presented above. However, in current market conditions having such a scheme requires continual interface with the various service providers. Such interface requires time and money. Therefore, having the interface and managing the interface with multiple SPs may reduce the attractiveness of this utility optimization scheme.

2) The Maximal Utility Model

Translating the problem into Minimum cost problem, the research assumes that 28.76 utility score translates to $ X. The customer must contact at least one SP for making any purchase at all. But for the case study assumes that the customer must contact the other two SPs to establish the purchases and track the transactions. Assuming a monthly cost per SP per service of $ 30.00 for the administrative work of ordering, tracking and payment management yields:

Cost of main SP1 (with 2 service of SP2 and 7 services of SP3): X+(7+2)*$30 = $X+$270

Cost of main SP2 (with 2 service of SP1 and 7 services of SP3): X+(7+3)*$30 = $X+$300

Cost of main SP3 (with 2 service of SP2 and 3 services of SP1): X+(3+2)*$30 = $X+$150

Thus, main **SP3** is chosen with minimal total monthly expenses of**: $X+$150**),

If the interface fee is larger than the difference in service price between the alternative provider and the main SP, buying the cheap item from the other SP with the additional cost would be more expensive than buying it from the main SP.

For example, suppose that $30 is equivalent to 0.3 units of utility. In that case, checking the services where other SPs have better utility than SP3 yields:

“Cloud search (CS)”: U(SP1, CS)-U(SP3, CS) =: 2.32-2.15 = 0.17<**0.3** → choosing **SP3** instead of SP1

ERP: U(SP1,ERP)-USP3,ERP) = 2.47-2.19 = 0.28<**0.3** → choosing **SP3** instead of SP1

Operating system (OS): U(SP2,OS)-U(SP3,OS)=2.44-1.68=0.76>**0.3** → Choosing **SP2**

Memory (M): U(SP2,M)-U(SP3,M) = 2.27-1.39=0.88>**0.3**→ Choosing **SP2**

Developer support (DS): U(SP1,DS)-U(SP3,DS) = 2.63-2.19=0.44>**0.3**→ Choosing **SP1**

Thus, the interface cost is: 3($30) = $90 and the overall utility is now: 28.76-0.17-0.28-0.3-0.3-0.3 = 28.76-1.35=27.41

It is obvious that as the interface cost goes up the optimal solution contains less and less such interfaces thus preferring a sole SP (The provider’s full bundle model).

3) The Direct Optimal Utility Comparison Approach

Let's assume that each interface between two services managed by two SP's reduces utility by 0.05 units. Following a computation of total utility in case SP1 in chosen as main cloud SP. The customer chooses each service according to maximal utility, reducing total utility for all interfacing services.

The model assumes that SP1 is the main service provider, and that the interfacing fee equals 0.05 utility units.

Thus, SP1 as main provider produces 28.31 utility units. Reducing utility by using interfacing fees does not produce a higher utility compared to using a model without fees.

Suppose now SP2 is the main provider, using other SP's services and paying them interfacing fees amounting 0.05 utility units. Following, the research presents the calculations of total utility.

TABLE VI. DIRECT OPTIMIZED UTILITY MODEL FOR MAIN SP1

SaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP2 – Fee	SP3 - Fee	Max (SP1, SP2-fee, SP3-fee)
	Data transfer	1.78	2.06	2.37	2.01	2.32	2.32
	Email services	2.04	1.76	2.84	1.71	2.79	2.79
	Cloud search	2.32	1.71	2.15	1.66	2.10	2.32
	Documents Mgt.	1.64	1.89	2.23	1.84	2.18	2.18
	ERP	2.47	1.30	2.19	1.25	2.14	2.47
PaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP2 – Fee	SP3 - Fee	Max (SP1, SP2-fee, SP3-fee)
	Operating system	1.38	2.44	1.68	2.39	1.63	2.39
	Memory	1.73	2.27	1.39	2.22	1.34	2.22
	Instance storage	1.30	1.89	2.23	1.84	2.18	2.18
	Developer support	2.63	1.82	2.19	1.77	2.14	2.63
IaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP2 – Fee	SP3 - Fee	Max (SP1, SP2-fee, SP3-fee)
	Relational Database services	1.64	2.06	2.37	2.01	2.32	2.32
	Storage standard vol.	1.45	2.44	1.94	2.39	1.89	2.39
	Backup	1.66	1.71	2.15	1.66	2.10	2.10
Total							**28.31**

TABLE VII. DIRECT OPTIMIZED UTILITY MODEL FOR MAIN SP2

SaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1 – Fee	SP3 - Fee	Max (SP1-fee, SP2, SP3-fee)
	Data transfer	1.78	2.06	2.37	1.73	2.32	2.32
	Email services	2.04	1.76	2.84	1.99	2.79	2.79
	Cloud search	2.32	1.71	2.15	2.27	2.10	2.27
	Documents Mgt.	1.64	1.89	2.23	1.59	2.18	2.18
	ERP	2.47	1.30	2.19	2.42	2.14	2.42
PaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1 – Fee	SP3 - Fee	Max (SP1, SP2-fee, SP3-fee)
	Operating system	1.38	2.44	1.68	1.33	1.63	2.44
	Memory	1.73	2.27	1.39	1.68	1.34	2.27
	Instance storage	1.30	1.89	2.23	1.25	2.18	2.18
	Developer support	2.63	1.82	2.19	2.58	2.14	2.58
IaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1– Fee	SP3 - Fee	Max (SP1, SP2-fee, SP3-fee)
	Relational Database services	1.64	2.06	2.37	1.59	2.32	2.32
	Storage standard vol.	1.45	2.44	1.94	1.40	1.89	2.44
	Backup	1.66	1.71	2.15	1.61	2.10	2.10
Total							**28.31**

Thus, SP2 as main provider produces 28.31 utility units same as SP1 as main provider. Reducing utility by using interfacing fees does not produce a higher utility compared to 28.76 using the pricing simple model without fees.

Suppose now SP3 is the main provider, using other SP's services and paying them interfacing fees amounting 0.05 utility units. Following, the research presents calculations of total utility.

TABLE VIII. DIRECT OPTIMIZED UTILITY MODEL FOR MAIN SP3

SaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1 – Fee	SP2 - Fee	Max (SP1-fee, SP2-fee, SP3)
	Data transfer	1.78	2.06	2.37	1.73	2.01	2.37
	Email services	2.04	1.76	2.84	1.99	1.71	2.84
	Cloud search	2.32	1.71	2.15	2.27	1.66	2.27
	Documents Mgt.	1.64	1.89	2.23	1.59	1.84	2.23
	ERP	2.47	1.30	2.19	2.42	1.25	2.42
PaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1 – Fee	SP2 – Fee	Max (SP1-fee, SP2-fee, SP3)
	Operating system	1.38	2.44	1.68	1.33	2.39	2.39
	Memory	1.73	2.27	1.39	1.68	2.22	2.22
	Instance storage	1.30	1.89	2.23	1.25	1.84	2.23
	Developer support	2.63	1.82	2.19	2.58	1.77	2.58
IaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1– Fee	SP2 – Fee	Max (SP1-fee, SP2-fee, SP3)
	Relational Database services	1.64	2.06	2.37	1.59	2.01	2.37
	Storage standard vol.	1.45	2.44	1.94	1.40	2.39	2.39
	Backup	1.66	1.71	2.15	1.61	1.66	2.15
Total							**28.46**

Thus, **SP3** as main provider produces **28.46** utility units, improving utility over SP1 and SP2 (which is 28.31).

Note that, the optimized pricing model using interfacing fees worth 0.05 utility units does not produce a higher utility compared to **28.76-5(0.05)=28.51** using the minimal cost approach of section 3.2.

V. DISCUSSION

Following, the research presents in Fig. III a sensitivity analysis assuming that SP3 is the main provider, computing the impact of varying interfacing fees on the calculations of best utility, according to all three models.

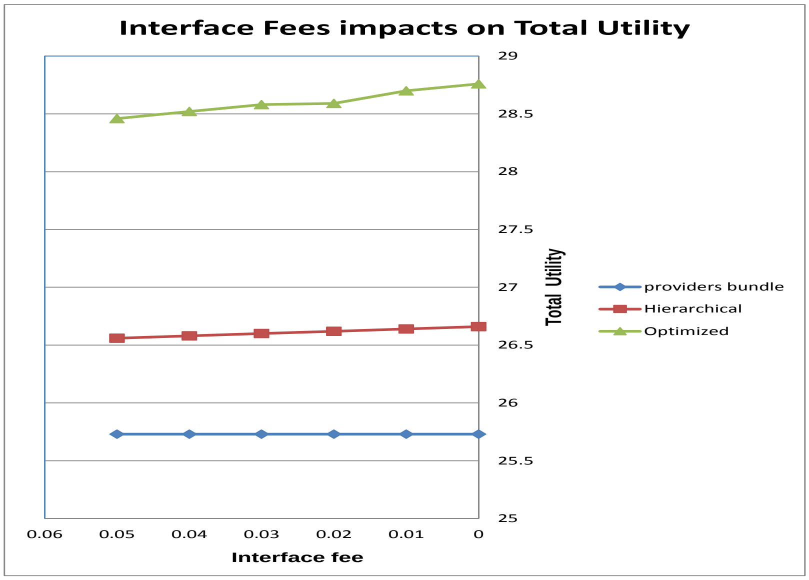


Fig. 3. Sensitivity Analysis assuming sp3 main Service Provider

First, the research assumes the interfacing fee reduces utility by 0.01 units. The research compares the calculated utility to the best utility produced by SP3. Table VIII presents computed utilities assuming an interfacing fee cost of 0.01 utility units (total utility: 28.7). Changing the assumption to interfacing fee cost of 0.02 units maximal utility provides 28.59 utility units.

Assuming the interfacing fee reduces utility by 0.03 units maximal utility provides 28.58 utility units.

Assuming the interfacing fee reduces utility by 0.04 units maximal utility provides 28.52 utility units.

Thus, a situation where all providers use standard interfaces in a free competition, consumers gain maximum utility. As interface costs of non-standard providers grow, the total utility for the consumer declines.

Notice that the providers-bundle model produces minimal utility, which is naturally unaffected by interface fees, since there are no interfaces.

The hierarchical model produces an improved utility over the bundle. Maximal utility is gained when no interface fees are used. Usage of interfaces diminishes utility since the consumer has to pay extra fees interfacing variety of providers.

Best utility is achieved using the optimized model. Interfacing fees impact on diminishing utility. Such expenses are needed as long as providers are using non-standard protocols, forcing consumers to pay for communicating between different standards. In a future situation when free market competition will force providers use standard protocols, consumers will be able to gain the maximal utility.

TABLE IX. UTILITY CALCULATIONS ASSUMING INTERFACE FEE COST OF 0.01 UTILITY UNITS

SaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1 – Fee	SP2 - Fee	Max (SP1-fee, SP2-fee, SP3)
	Data transfer	1.78	2.06	2.37	1.77	2.05	2.37
	Email services	2.04	1.76	2.84	2.03	1.75	2.84
	Cloud search	2.32	1.71	2.15	2.31	1.70	2.31
	Documents Mgt.	1.64	1.89	2.23	1.63	1.88	2.23
	ERP	2.47	1.30	2.19	2.46	1.29	2.46
PaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1 – Fee	SP2 – Fee	Max (SP1-fee, SP2-fee, SP3)
	Operating system	1.38	2.44	1.68	1.37	2.43	2.43
	Memory	1.73	2.27	1.39	1.72	2.26	2.26
	Instance storage	1.30	1.89	2.23	1.29	1.88	2.23
	Developer support	2.63	1.82	2.19	2.62	1.81	2.62
IaaS	Service name	SP1 Utility	Sp2 Utility	SP3 Utility	SP1– Fee	SP2 – Fee	Max (SP1-fee, SP2-fee, SP3)
	Relational Database services	1.64	2.06	2.37	1.63	2.05	2.37
	Storage standard vol.	1.45	2.44	1.94	1.44	2.43	2.43
	Backup	1.66	1.71	2.15	1.65	1.70	2.15
Total							**28.70**

VI. Conclusions

This paper presents a model that is based on conjoint analysis method to measure the utility of CC service utilities under the major stages in the scale between full bundling of CC services, through partial bundling, to free market conditions. The customer is trying to maximize his/her overall utility while facing open tariffs of various services from the various SPs.

Current SPs practices of bundling services blocks and obstructs market competition in cloud computing. In the long run (with the addition of SPs) economic competition theory predicts that full bundling will disappear with the rise of free market forces. As CC service competition will develop, and tools to enable this will be more common consumers are bound to look for an optimized combination of services and service providers that maximize their utility under the prevalent market conditions. The research presented three major stages of shifting to a free market, and showed the optimal customers' strategy in each stage. At first, the research ignores SP interface/monitoring costs, and shows that as the level of freedom to switch services grows, so is the overall utility. However, the cost of interfacing multiple SPs is a tradeoff, very large interface cost have the same effect as bundling, and are making the single SPs more attractive. Finally, the research presents a detailed case study implementation of the model and its stages and strategies illustrate the advantages of the proposed strategies compared to existing practices used by cloud computing consumers.

Three directions are identified for future research, based on this paper:

1) Examining the effects of uncertainty on consumer’s choice.

2) Considering risk and risk aversion on consumer’s choice.

3) Developing a model which describes the behavior of both the consumers and the CC service providers based on game theory.

References

[1] T. Aoyama, and H. Sakai, “Inter-Cloud Computing”, Business Information Systems Engineering, Vol. 3, 2013.

[2] M. A. Armbrust, R.Fox, A.J. Griffith, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia., Above the Clouds: A “Berkeley View of Cloud Computing”. Technical Report, Berkeley:CA,.2009.

[3] K. Bogataj, and A. Pucihar, “Business Model Factors Influencing Cloud Computing Adoption: Differences in Opinion”, BLED 2013 Conference Proceedings, Bled:Slovenia, 2013.

[4] Z. Chen, F. Han, J. Cao, X. Jiang, and S. Chen, “Cloud Computing-Based Forensic Analysis for Collaborative Network Security Management System”. Tsinghua science and technology, Vol 18, No. (1, 2), 2013.

[5] A. Gill, D. Banker, and P. Seltsika, “Moving Forward: Emerging Themes in Financial Services Technologies Adoption”, Communications of the Association for Information Systems: Vol. 36, Article 12, 2015.

[6] P. E. Green, A. M. Krieger, and Y. J. Wind, “Thirty Years of Conjoint Analysis: Reflections and Prospects”. Interfaces Vol. 31, No.3, 2001, pp. 56-73.

[7] A. Hollobaugh, “Hosting.com Cloud Computing Trends Repor”t, Technical Report, 2009. http://www.hosting.com.

[8] P. Koehler, A. Anandasivam, M. Dan, and C. Weinhardt, “Customer heterogeneity and tariff biases in cloud computing”. Thirty First International Conference on Information Systems, St. Louis 2010 1 (ICIS 2010) proceedings, 2010.

[9] P. Koehler, A. Anandasivam, and A. Dan, “Cloud services from a consumer perspective”, AMCIS 2010 Proceedings, 2010.

[10] W. F. Kuhfeld, “Marketing Research Methods in SAS Experimental Design, Choice”, Conjoint and Graphical Techniques. Technical Report, 2009.

[11] J. J. Louviere and G. G. Woodworth, “Design and Analysis of Simulated Consumer Choice or Allocation Experiments: an Approach based on Aggregate Data”. Journal of Marketing Research, Vol. 20, 1983, pp. 350-367.

[12] Y. Mansouri, A. N. Toosi, and R. Buyya, “Brokering Algorithms for Optimizing the Availability and Cost of Cloud Storage Services”, 2013 IEEE International Conference on Cloud Computing Technology and Science, 2013.

[13] P. Mell, and T. Grance, “The NIST definition of cloud computing”, National Institute of Standards and Technology, NIST, Vol. 53 No. 6, 2009, p. 50.

[14] F. Paraiso, N. Haderer, P. Merle, R. Rouvoy, and L.Seinturier, “A Federated Multi-Cloud PaaS Infrastructure”, 2012 IEEE Fifth International Conference on Cloud Computing, 2012.

[15] T. Pueschel, A. Anandasivam, S. Buschek, and D. Neumann, “Making money with clouds: Revenue optimization through automated policy decisions”. 17th ECIS - European Conference on Information Systems, 2009.

[16] U. Z. Rehman, F. K. Hussain, and O. K. Hussain, “Towards Multi-Criteria Cloud Service Selection”, 2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011.

[17] J. S. Forrester, “Cloud predictions for 2014: Cloud joins the IT portfolio”, http://blogs.forrester.com/james_staten/13-12-04-cloud_computing_predictions_for_2014_cloud_joins_the_formal_it_portfolio, accessed 02 March 2014.

[18] L. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, “A Break in the Clouds: Towards a Cloud Definition”, Editorial note. ACM SIGCOMM (2009). Computer Communication Review 50, Volume 39, Number 1, 2009.

[19] A. Velte, R. Elsenpeter, and T. J.Velte, “Cloud Computing: A practical approach”. Tata McGraw-Hill Education Pvt. Ltd, 2009.

[20] W .Venters, and E.A.Whitley, “A critical review of cloud computing: researching desires and realities”, Journal of Information Technology, Vol. 27, No.3, 2012, pp. 179-197.

[21] R. Weiber, and D. Mühlhaus, “Auswahl von Eigenschaften und Ausprägungen bei der Conjointanalyse”. In “Conjointanalyse”, by D. Baier and M. Brusch, Heidelberg: Springer, 2009, pp.43-58..

[22] C. Weinhardt, B.Blau, and J. Stößer, “Cloud Computing – A Classification, Business Models, and Research Directions”. Business & Information Systems Engineering, 2009.

[23] E. Weintraub, and Y. Cohen, “Cost Optimization of Cloud Computing Services in a Networked Environment”, (IJACSA) International Journal of Advanced Computer Science and Applications ,Vol. 6, No. 4, 2015, pp. 148-157.

[24] Q. Zhang, L. Cheng, and R. Bautaba, “Cloud computing: State-of-the-art and Research challenges”. Journal of Internetional Service Applications, Vol. 1, 2010, pp. 7-18.

Design and Realization of Mongolian Syntactic Retrieval System Based on Dependency Treebank

S.Loglo
Language Research Institute
College of Mongolian Studies
Inner Mongolia University
Huhhot, Inner Mongolia Autonomous Region, China

Sarula
Department of Journalism and publishing
College of Mongolian Studies
Inner Mongolia University
Huhhot, Inner Mongolia Autonomous Region, China

Abstract—**In the past seven years, Language Research Institute of Inner Mongolia University has constructed a 500,000-word scale Mongolian dependency treebank. The syntactic treebank provides a favorable data platform for language research and information processing. In order to effectively use the treebank, we have designed and implemented a graphical syntactic information retrieval system based on the Mongolian dependency treebank. As an application system, this retrieval system offers search and statistical analysis on word, phrase, syntactic fragment and syntactic structure level.**

Keywords—Mongolian Language; Dependency Grammar; Dependency Treebank; Syntactic Retrieval; Information Retrieval

I. Introduction

Language Research Institute of Inner Mongolia University has constructed a 1-million-word modern Mongolian corpus in a span of eight years from 1984 to 1991 and expanded it twice into what is now a 10-million-word corpus. The 1-million-word corpus contains materials from novels (19.6%), textbooks (50.3%), newspapers (9.8%) and politics (22.9%) [1]. In corpus annotation, the 10-million word corpus has been completed the part-of-speech tagging [2][3] and fixed phrase tagging [4]. And some shallow parsing is carried out on the 1-million word corpus, such as phrase tagging [5][6][7], automatic sentence segmentation [8] and automatic predicate segment recognition [9].

From 2008 to 2011, funded by National Social Science Foundation and National Natural Science Foundation, using the method of automatic parsing and manual proofreading, Language Research Institute of Inner Mongolia University has constructed a 500,000-word Mongolian dependency treebank (MDTB) based on middle school Mongolian textbooks that were extracted from the 1-million-word modern Mongolian corpus [10]. MDTB has an annotation set of 17 dependency relations under 5 categories [11]. The 5 categories are special relation, dominant relation, conjunctional relation, auxiliary relation and non-syntactical elements. The 17 dependency relations include: key word in a sentence (HEAD), independent element (INDE), subject (SUBJ), direct object (DOBJ), indirect object (IOBJ), attribute (ATT), adverbial (ADV), coordinate (COO), appositive (APP), summarization (SUM), time-local words-auxiliary (TL-AUX), postposition-auxiliary (PP-AUX), modal particles-auxiliary (MP-AUX), modals-auxiliary (M-AUX), auxiliary verbs-auxiliary (AV-AUX), contact verb-auxiliary (CV-AUX) and conjunction-auxiliary (CJ-AUX). In the form of annotation, MDTB uses two types of labeling, namely the brackets annotation and graphical annotations. This treebank contains 461,240 words in 31,722 sentences. The average sentence length is 14.54 words.

Mongolian dependency treebank contains rich syntactic information, so the researchers can obtain all kinds of information about syntax. On the dependency treebank, researchers also can perform statistical analysis and example sentence extraction. Therefore, it provides convenience for the study and research of Mongolian traditional linguistics and computational linguistics [12] [13]. However, at present, treebank is usually used as a training and evaluation data for syntactic parsing [14], but research about the syntactic information retrieval is few and far between. This paper is designed to expound a syntactic treebank retrieval system based on the application system of the dependency treebank. The retrieval functions allow researchers to do enquiry and statistical analysis on word, phrase, sentence constituents, syntactic fragment and syntactic structure.

II. Design and Realization of Mongolian Syntactic Retrieval System

Dependency tree-based Mongolian syntactic retrieval system is divided into two parts, syntactic tree display module and retrieval statistics module. Realization of the two modules is presented as follows.

A. Display Module of the Syntactic Tree

Graphic display lies in the heart of treebank operation as an essential technology. However formatted a treebank is, in text or in graph, the output module can draw a complete tree for each sentence, as shown in the Fig.5. The left window displays corpus texts of which the current sentence is displayed in selected model. The right window displays the syntactic tree of the current sentence.

Fig.1 shows the node structure of the syntactic tree. Each node on the syntactic tree may have n child nodes which are arranged from left to right according to the sequence in which dependency relations were established. But this order dampens the readability of treebank. To recover the original order of brother nodes, we have added sorting function to the output model. The output module also provides multiple optional display modes which are presented as follows.

Sponsored by National Natural Science Foundation (60763033), National Natural Science Foundation (61262046), and National Social Science Foundation(10CYY022).

1) Open or close lexical information display function;

2) Contract or expand descendant nodes; and

3) Display the entire syntactic tree or only display the search results.

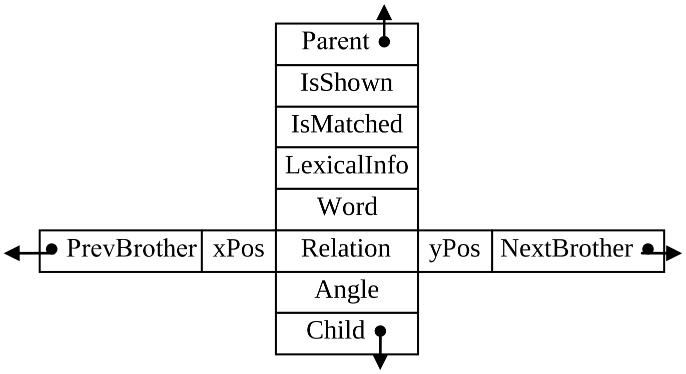


Fig. 1. Node structure of a dependency tree

In Fig.1, ***PrevBrother*** stores a pointer that points to the node's previous brother, ***NextBrother*** stores a pointer that points to the node's next brother, ***Parent*** stores a pointer that points to the node's parent node, ***Child*** stores a pointer that points to the node's child node, ***Word*** stores the node's word, ***LexicalInfo*** stores the node's lexical information, ***Relation*** stores the node's dependency relation, ***xPos*** and ***yPos*** stores the node's horizontal and vertical ordinate, Angle stores the node's inclined angle of dependency arc, ***IsShown*** denotes whether the node's descendent nodes are shown or not, and ***IsMatched*** denotes that the node is among the research results.

The output algorithm is as follows:

```
VOID ShowTree (CTree *T,int ShowMode)
//T denotes the dependency tree;
//ShowMode==”0” indicates that the program will
//display all syntax trees; and "ShowMode==1" indicates
//that the program will display the results of the search.
{If (ShowMode==1 &&
T->bSearchResult==FALSE) return;
//Visit RootNode on the dependency tree T;
If(!RootNode->IsShrunk())
// denotes that the root node is not shrunk;
ShowNode(RootNode);//draw the node (RootNode)
If(RootNode->bShowLexcicalInfo==TRUE)
ShowLexicalInfo(RootNode);
//shows the node’s lexical information;
SortChildrens(RootNode);
//Sort the child nodes of RootNode;
//Traverse subtree forest of root nodes pre-orderly;
For(i=0;i<n;i++)a
//n is the number of child nodes of RootNode
ShowTree (CTi);
//CTi denotes a sub-tree whose root node is the ith
//child of T
Return;
}
```

Algorithm1. Dependency tree display algorithm

B. Design of Mongolian Syntactic Retrieval Algorithm

Treebank is an important resource for syntactic analysis and evaluation, word sense disambiguation and semantic analysis. MDTB provides a favorable data platform for Mongolian language research and information processing. At present, the use of treebank is mainly achieved through a variety of retrieve technology-based statistical methods. As such develop an efficient search algorithm is very necessary for treebank-based systems [15] [16].

The dependency treebank herein adopts two different storage formats, text and graph. Text format is for treebank that targets all users and can be opened and edited by any text editing software. Graphical format, which MDTB adopts, benefits both output and retrieval, although it does not perform better than text format in terms of space utilization. Based on graphical storage format, we have designed a treebank retrieval algorithm with sub-tree query function. The query conditions can be a sub-tree, a word or a syntactic fragment with n nodes. Each node can have one or multiple characteristic values such as vocabulary (can use wildcards like ‘*’ and ‘?’), parts of speech, sub-categorization, morphology, dependency relation type, father node and child node.

The retrieval algorithm is as follows:

1) Traverse syntactic tree pre-orderly to find root node (sRoot) of query condition;

2) If a node (tNode) in the dependency tree satisfies the requirement of the query condition’s root node (sRoot), then, to find all the child nodes of sRoot in the child nodes of the tNode;

3) If all of sRoot’s child nodes are found in step (2), recursively call step (2) until nodes that meet the requirements can no longer be found or the rightmost descendent nodes of query conditions are found;

4) Continue to find sub-trees among the remaining nodes of the current tree (excluding traversed nodes); and

5) Treebank search needs to call step (1) to step (4) repeatedly.

tNode represents one particular node on syntactic tree, and sRoot denotes root node of one sub-tree under given query conditions at given moment (including query condition itself).

C. A Syntactic Retrieval Example

The rationale of retrieval algorithm is explained by finding juxtaposed attributive in the following sentence.

Fig.2 shows the dependency tree of the following sentence,

bi uran=sibauxay-yin xatagu=sirgagu ajilči=xödelműriči ǰorig=sanag_a bolon uran=narin egűr sűlǰixű mergeǰil-i űnen=sedxil-eče-ben bisiren_e. (I admire from the bottom of my heart the brave and hard-working bird’s strong will and superb nesting skill).

The dotted lines represent syntactic fragments that meet the requirements of query conditions.

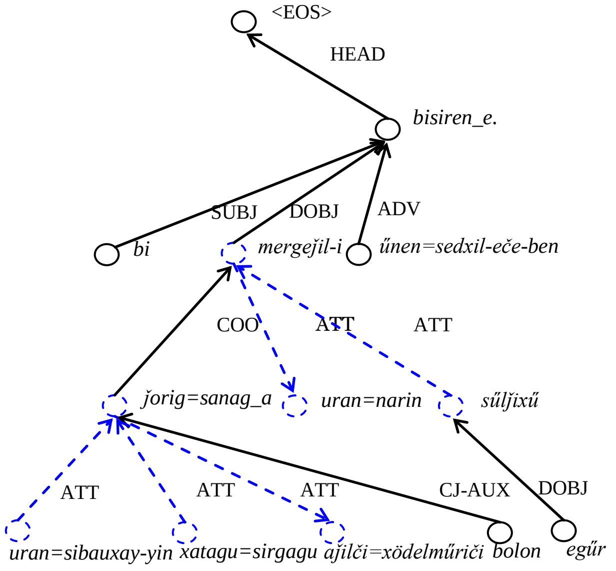


Fig. 2. Dependency Tree of the example sentence

Query condition can be a word or phrase node, a dependency arc or a syntactic fragment of any size. In the query condition, node can have 16 attribute values including the word or phrase itself, part of speech, syntactic relation and affix. In the process of search, a query interface as shown in the Fig.5 will pop up. A new node will be added by clicking the white dot. A dependency relation can be established between two nodes by dragging the mouse. Fig.3 shows the query conditions of this example.

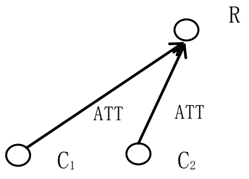


Fig. 3. Query Condition

1) Find nodes that match with R in dependency tree using pre-order traversal. As R itself has no constraints, every node in the dependency tree meets its requirement. The key is to check whether the node has two ***ATT*** *child nodes. In the process of traversal, the node "mergeǰil-i" meets the requirement, as shown in Fig.4 (a).*

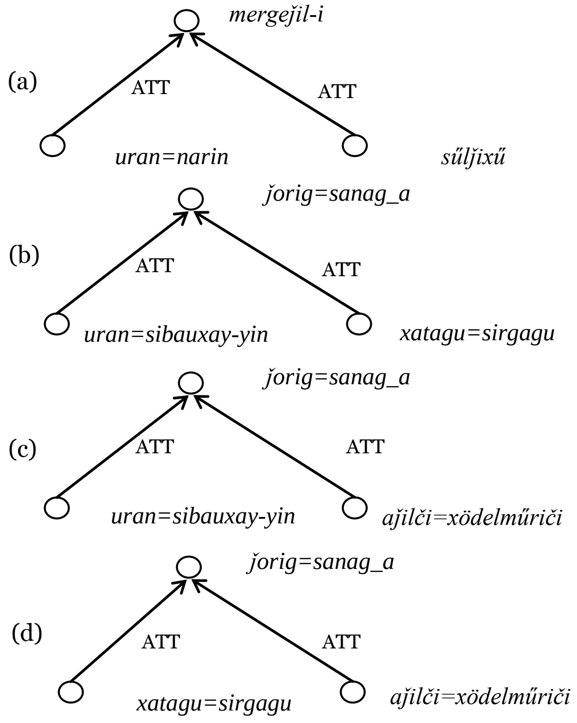


Fig. 4. Node combinations that meet the query condition

As the enquiry condition only contains two-layer nodes, there is no need for recursive query.

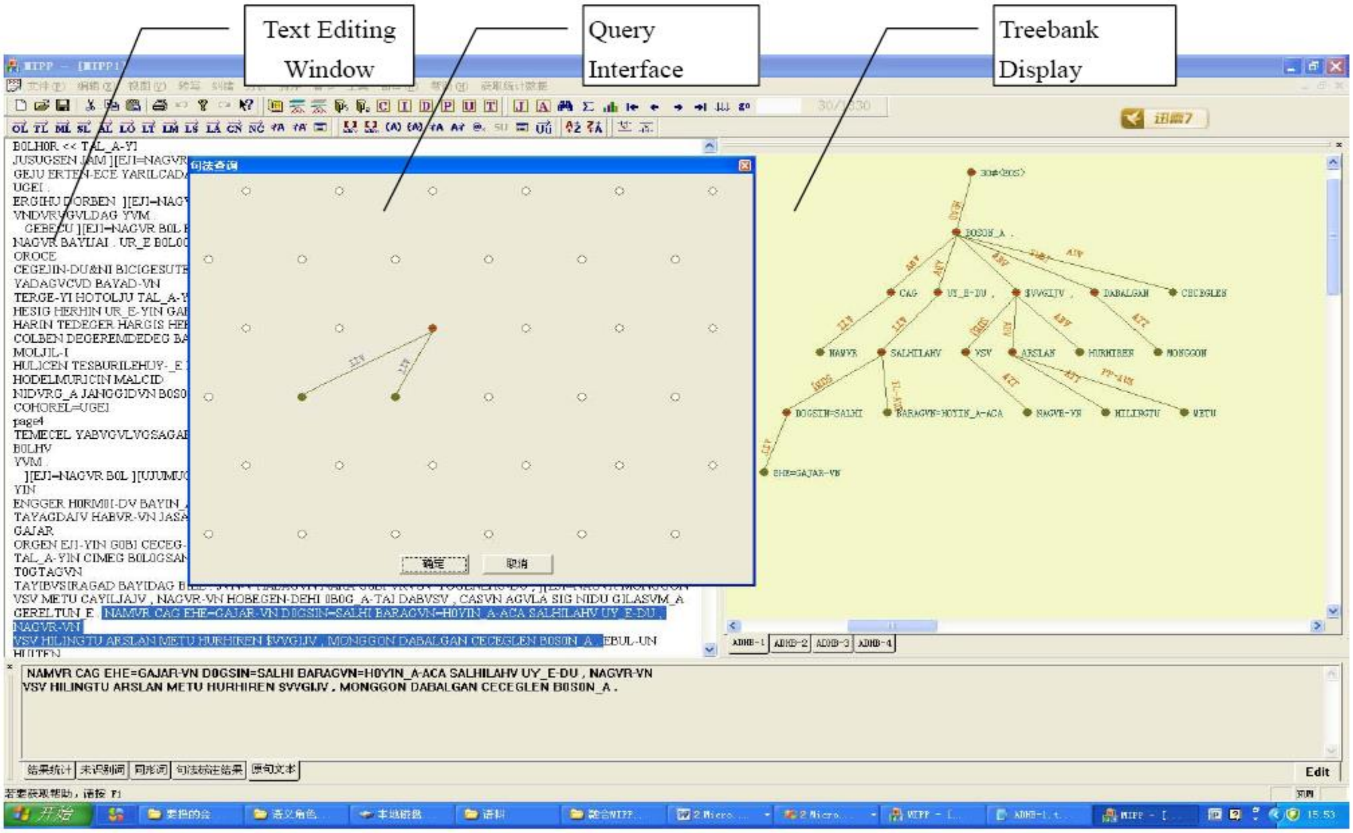
2) Label with different colors syntactic fragments that have been found. Continue to traverse the dependency tree to find the next eligible syntactic fragment.

3) The combinations where node " ǰorig=sanag_a" and its child nodes meet the requirement are as shown in Fig.4 (b)—(d).

4) Search is done when the remaining nodes in the dependency tree have been traversed and no eligible fragments are found.

It is worth noting that the program will restore the treebank and clear the traces left from the previous query operation before next query. If the search results need to be preserved, a copy needs to be saved by using the pertinent functions in this program.

Statistical analysis of syntactic fragment is done based on query. Each search provides relevant statistical data, including the number of times a syntactic fragment appears and the number of sentences that contain the syntactic fragment.



NAMVR CAG EHE=GAJAR-VN DOGSIN=SALHI BARAGVN=HOYIN_A-ACA SALHILAHV UY_E-DU , NAGVR-VN
VSV HILINGTU ARSLAN METU HURHIREN SVVGIJV . MONGGON DABALGAN CECEGLEN BOSON_A .

Fig. 5. The Editing, showing or query interface of Mongolian dependency Treebank

III. Conclusion

The dependency tree that this Mongolian syntactic retrieval system is based differs from binary phrase structure tree in terms of node types and tree structures. Such difference cannot be effectively handled at the current stage. Moving forward, we will add to this system display, editing and retrieval function for phrase structure tree and bidirectional conversion between the two kind of tree structures. This system has good universality, and has no particular relationship with language per se. As such it can be applied to other languages' treebank for editing and retrieval operations.

Reference

[1] Language Research Institute of Inner Mongolia University, "About modern Mongolian corpus", Journal of Inner Mongolia University (Humanities & Social Sciences), 1992, vol.24, N0.1, pp. 1–5.

[2] HUA Shabao, "AYIMAG– A POS tagging system for Mongolian corpus", Journal of Inner Mongolia University (Humanities & Social Sciences), 1999, vol.31, N0.5, pp.31–35.

[3] Zhang Guanhong, S.Loglo and Odbal, "Fusion of morphological features for Mongolian part of speech based on maximum entropy model", Journal of Computer Research and Development, 2011, vol.48, N0.12, pp.2385–2390.

[4] S.Loglo and Sarula, "Research on Mongolian lexical analyzer based on NFA", Proceedings of 2010 IEEE International Conference on Intelligent Computing and Intelligent Sytems, Xiamen, China, 2010, vol.2, pp.240–245.

[5] Hua Shabao and Dabhurbayar, "A Phrase-tagging Research in Mongolian Corpus", Journal of the Central University for Nationalities (Philosophy and Social Sciences Edition), 2006, vol.33, No.5, pp.64–67.

[6] Hua Shabao, "A tagging strategy of Mongolian phrases", Journal of the Central University for Nationalities (Philosophy and Social Sciences Edition), 2003, vol.30, N0.5, pp.98–100.

[7] Wulan, Dabhurbayar, GUAN Xiaoda and ZHOU Qiang, "Phrase structure parsing of Mongolian", Journal of Chinese Information Processing, 2014, vol.28, No.5, pp.162–169.

[8] Wang Serguleng, "Rule-based Mongolian sentence automatic segmentation," Journal of Inner Mongolia University (Philosophy & Social Sciences (Mongolian Edition)), 2009, vol.38, No.3, pp.51–55.

[9] Wang Serguleng, D.Sarana and Nasunurtu, "Design and realization of automatic annotation for modern Mongolian predicate segment", Proceedings of 11th national symposium on minority languages, Xishuangbanna, China, 2007, pp.420–427.

[10] S.Loglo and Sarula, "Construction of a Mongolian dependency treebank", International Journal of Knowledge and Language Processing, 2014, vol.5, No.2, pp.32–42.

[11] S.Loglo, "Research on the modern Mongolian syntactic tagging system based on the dependency grammar", Mongolian Studies of China, 2011, vol.39, No.2, pp.116–119.

[12] Liu Haitao, "Dependency grammer (from Theory to practice)", Science Press, Beijing, China, 2009, pp.1–15.

[13] Jiří Mírovský, Netgraph, "A tool for searching in prague dependency treebank 2.0", Proceedings of the TLT 2006, pp.211–222.

[14] S.Loglo and Sarula, "A rule-based Mongolian dependency parsing model", International Journal of Knowledge and Language Processing, 2013, vol.4, No.3, pp.27–37.

[15] Jiří Mírovský, "Searching in the prague dependency treebank", Published by Institute of Formal and Applied Linguistics, Czech Republic, 2009.

[16] Laura Kallmeyer, "On the complexity of queries for structurally annotated linguistic data", Proceedings of ACIDCA, 2000, pp.1–6.

Comparison of Burden on Youth in Communicating with Elderly using Images Versus Photographs

Miyuki Iwamoto
Graduate School of Engineering and Science, Kyoto Institute of Technology
Kyoto, Japan

Noriaki Kuwahara
Graduate School of Engineering and Science, Kyoto Institute of Technology
Kyoto, Japan

Kazunari Morimoto
Graduate School of Engineering and Science, Kyoto Institute of Technology
Kyoto, Japan

***Abstract*—Conversation is a good preventative against behavioral problems in the elderly. However, caregivers are usually very busy tending to patients and lack the time to communicate extensively with them. Toward overcoming such problems actively listening volunteers have more opportunities to communicate with the elderly, but the number of skilled volunteers is limited. Therefore, we investigated conversational support systems for inexperienced volunteers; such systems usually include content such as photographs, videos, and music. We expected that the volunteers would feel less stress when using videos instead of photographs for conversational support because the former provided both volunteers and patients with richer information than the latter. On the other hand, photographs gave patients more chances to talk with volunteers. However, there has been no research to date on the effect of content type upon stress and conversational quality. In this paper, we compared using photographs with using video from such viewpoints.**

Keywords—elderly; reminiscence videos; senior care home photographic image

I. Background and Purpose

Japanese society is recently facing the problem of having a "super-aging" population. The proportion of aged people is growing. The population over 65 years old, so-called "Baby-Boomers," is now over 30 million, and it is anticipated to be 36 million in 2030. [1]

The rate of families consisting of old couples and old singles is increasing. In some cases s/he may pass an entire day without speaking a word, which can lead to a disuse of cognitive functions and a heightened risk of dementia and/or depression.

We need to understand such physical and psychological characteristics in our communication with aged people. For example, we must respect the damage to their pride at losing their visual, auditory, and cognitive functions. Experts such as clinical psychologists, therapists, or listening volunteers may be able to deal with these issues in their communication efforts. However, their numbers are insufficient to meet the current needs.

Thus, the younger generations are expected to be talking partners for aged people, but there is a problem in that they are unfamiliar with how to communicate with the elderly because the vast majority/ many grew up in small families without grandfathers or grandmothers.

There have been some attempts to solve the problem using a picture or a video as a trigger for conversation. But there is no research regarding the mental burden felt by the partner students/ volunteers. That means the research has concentrated on communication support systems to improve QOL for aged people, while the mental burden of the young partners has been neglected. Some volunteers have said that in their experiences they felt less mental burden in watching a video than in looking at a photo. Yet the effects in quality and in quantity of conversation based on the contents of the support system have not been examined. It is therefore unclear whether there is a difference between how photos and videos affect their mental burden.

We have also used photographs and videos as supporting tools for communication between aged people with dementia and students/ volunteers, as shown in the following figure. We have noticed that although these materials make communication smoother and offer a valuable opportunity for the aged people, some degree of mental burden is imposed upon the students/ volunteers.

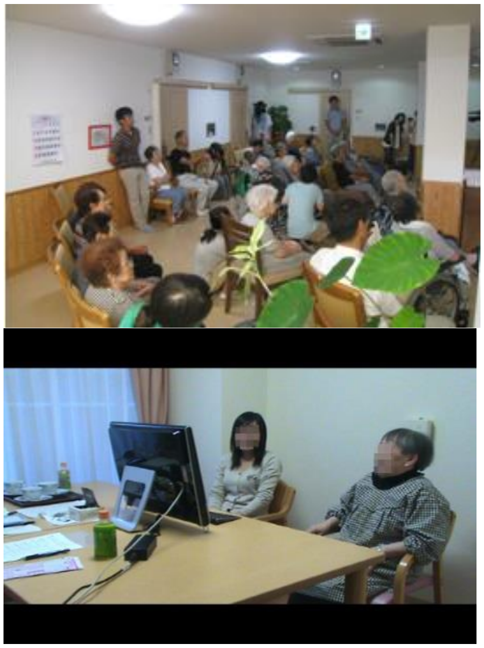


Fig. 1. Volunteers utilizing conversation support content

Many caregivers and volunteers untrained in listening often feel considerable mental burden. There is some research on promoting conversation between patients and caregivers or volunteers, or among patients, by providing topics and/ or sharing photos or videos. (Exam. [2][3][4]) This focused on what is known as reminiscence technique, which is effective for controlling dementia and reducing the mental burden placed upon the partner students/ volunteers. We call the information media which provide the topics (hereafter, "media") "communication support content."

The media in general include photos, videos, and music. Their effectiveness with regard to the mental burden felt by caregivers, or the quality of communication between patients and caregivers, is unknown.

We examined the differences in the mental burden and the quality of communication between patients and caregivers/ volunteers when they used photos versus when they used videos as communication support content in order to find the best medium for communication.

We measured the mental burden in reference to our research presented at the Human Interface symposium [6]. We compared whether less burden was felt in communication using photos or videos, improving upon problems found in the former research. As for quality and quantity of communication, we estimated these by the length and sight lines observed in the communicative exchange, an approach used in the 5-stage subjective assessment from previous conversational support studies. [2]

We found in the previous research that measuring the mental burden was difficult because of the individual characters of the aged people. In that project, a pair of students would face a patient with each having the chance to talk only once. We changed the situation in this case such that each student interviewed five patients. We could then evaluate the mental burden by averaging the influence of compatibility between the patient and the partner.

There was an additional problem concerning the accuracy of the length of conversation, as it could be interrupted by searching for a photo or video via the internet in the previous research. This was solved by preparing the photos and videos in advance.

The subjectivity of our questions and video recording was furthermore an issue. . Therefore we added measuring heart rates as an objective aspect this time.

Finally, the variation in content when students searched for photos or videos on the internet was also an issue. This could present difficulty in measuring the degree of mental burden because the photo or video may or may not have been interesting or favorable to them. In the case of the present research, we used the same content in order to control the results measuring the quality of the conversation.

We conducted an experiment to improve upon these points.

II. Experiment

A. Outline

We eliminated art-related topics because there was a great generation gap between the patients and students, and some topics were thought improper. We chose the topic categories of food, events, and plays from the former experiment. As for subjects, one student interviewed five patients and we estimated an average of the mental burden while excluding personal compatibilities.

We measured the length of conversation, preparing the photos and videos in advance to eliminate the burden of having to search for these during the conversation. We counted the number of heartbeats as an objective marker adding to the subjective estimation of mental burden.

The content of photos and videos in the categories of food, events, and plays were provided in advance. The photos were still pictures from the videos to examine the quality and quantity of the conversation and the related mental burden. Using the same content allowed us to control the influence of difference in contents.

The questions from the 5-stage subjective assessment were applied to estimate the quality of conversation after each experiment, as we had done before. We measured the quantity by the length of the communicative exchange in the video recording and looking at sight lines.

B. Evaluation items

We carried out checks on the degree of burden resulting from the conversation upon the students. A stress check board was positioned out of sight of the elderly patients (Fig.2). The numbers "1-7" were written on this stress check board, with "1" equating to no (stress) burden during the conversation, and "7" meaning that a great deal of stress was felt. At the beginning of the experiment, the students' stress levels were at "1". The students' stress levels were then measured each minute, at which time students could chose "1-7" according to the degree of stress they were experiencing during the conversation.

This represents the degree of stress felt in the conversation at a given time. During the conversation, the facial expressions and movements (appearance and nods) of the students and the elderly were recorded. We also measured the students' heart rates in order to determine their conversational stress levels.

During the evaluation of the conversational quality, a questionnaire was given once during each experiment according to the fifth stage subjective evaluation.

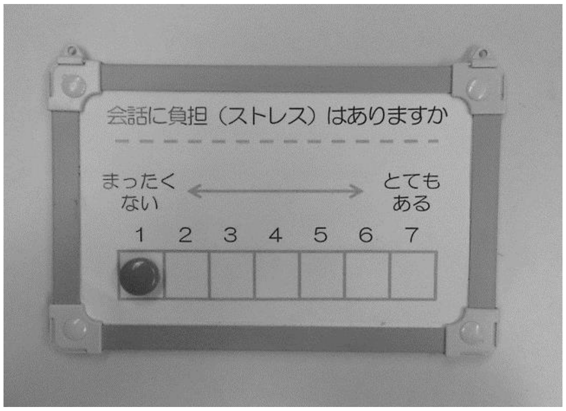


Fig. 2. Stress Check Board

C. Experiment environment

The layout of the experiment environment is shown in Figures3 and 4 below. We borrowed a nursing home room, in which we placed chairs side by side. We used a desktop PC to which we uploaded photos and videos fitting into the above-mentioned categories in order to facilitate the 10 minute-conversation. We asked the student to point to the appropriate number on the check board in a way hidden from the patient's eyes, preventing the student from drawing the patient's attention while they were sitting side by side.

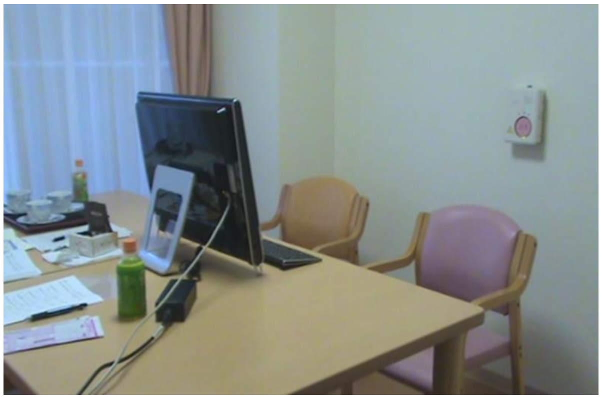


Fig. 3. Experiment environment

D. The Subject

The participants were 2 students, both of whom were 23 years old. Their degree of interpersonal skill was diagnosed in advance by the Yatabe-Guilford sociability personality diagnostic test. The patients were 8 senior ladies, ages 82 to 90, who were suffering from mild dementia.

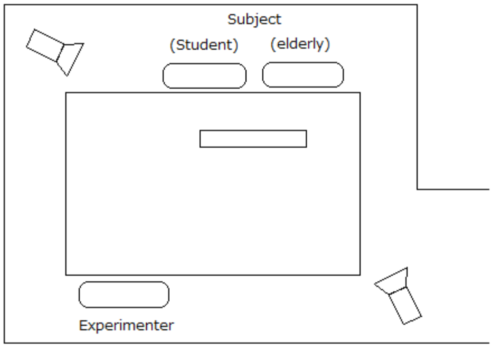


Fig. 4. The layout of the experimental environment

E. Procedure

The students and the patients talked while viewing a photo or a video through the PC. We prepared 20 photos each for the categories food, events, and plays, and these were shown through the flow of the conversation. We prepared a 10-minute video which was displayed on the PC monitor throughout the 10-minute conversation.

The student indicated the degree of mental burden on the stress check board each minute. The board was hidden from the patient's eye. Our purpose was to check the mental burden felt by the student during the conversation, not to check their complaints regarding the patient. Thus, the student was able to point out their feeling without having to be concerned about the patient. They could declare their feeling honestly. The board was sometimes shown in the middle of their conversation but they continued to talk at the moment they pointed. They answered the questions from the 5-stage subjective evaluation each time after the experiment. The heart rates of the participant students were also checked in each case to measure their mental burden. They placed the counters on their bodies in advance.

III. Result

Fig.5 shows the stress level of Student A and B (simply A or B below). The horizontal axis represents the photos and videos of subjects A and B, respectively, and the vertical axis represents the accumulation on the numerical stress check board. This is a comparison of the students' stress levels in the case of looking at photos and in the case of watching videos when they felt a sense of mental burden upon continuing the conversation using support contents.

From the diagnosis of characters, A showed 58% interpersonal skill and 50% sociability while B showed 75% and 61%, respectively. The students showed a difference in interpersonal skill. We compared the results of their mental burden from the conversations. A seemed to feel more mental burden when he showed photos to the patient.

We counted the heart rates of A and B and we took averages of the former half and the latter half in the 10-minute conversation with each subject. The vertical axis in Fig.6 and in Fig.7 show the number of heartbeats counted [number/ minute], and the horizontal axis show the experiment times in the former half and the latter half.

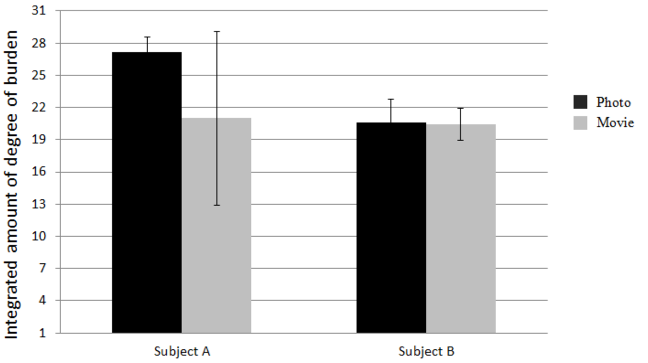


Fig. 5. Integration of the 10 minutes of degree of burden

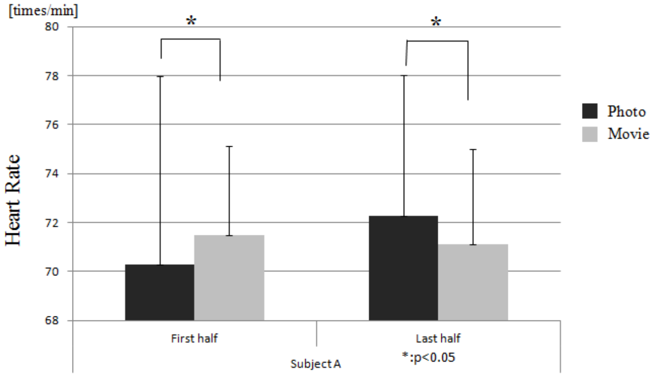


Fig. 6. The first half for Subject A, the average heart rate for the second half

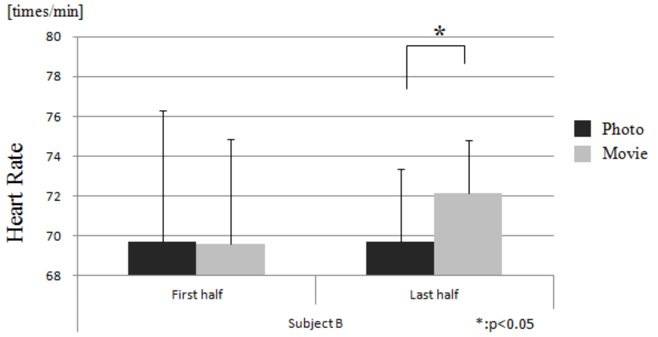


Fig. 7. The first half for Subject B, the average heart rate for the second half

We found no substantial change in the heart rates of A either in the former half or the latter half, as shown in Fig.6. In the former half, his heart rate was higher when showing videos, but in the latter half, it was higher when showing photos. In neither the former nor latter half was a significant difference in hear rate observed in the cases of photos or videos as a result of the (significance level 5%) t-test.

The heart rate of B had no difference at the beginning either in showing photos or in showing videos. In the latter half, his heart rate showed a greater change when showing photos than when showing videos. There was a significant difference by t-test in his heart rate between showing photos and showing videos.

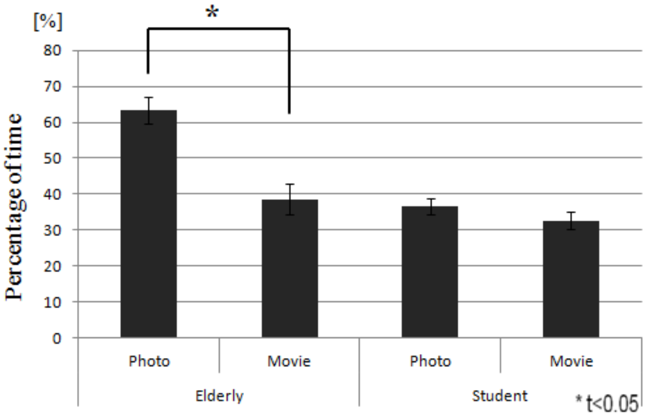


Fig. 8. Speech Time (Subject A)

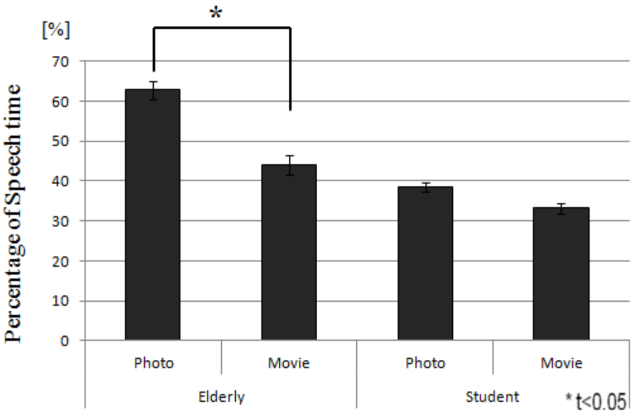


Fig. 9. Speech Time (Subject B)

Fig.8 (A) and Fig.9 (B) are the results of conversation time counted from the video recording when showing photos and when showing videos. The vertical axis is the rate (%) of the length of the communicative exchange over the 10 minutes, and the photos and the movies of the student and those of the patient. These figures show that either A or B, or the patient or the partner, talks longer when showing photos than when showing videos, though the difference in the time of the communicative exchange on the part of the patient was larger than that of the student. The results of the t-test (5% significance) were significantly different with regard to the length of the communicative exchange on the part of the patient and the student.

We also examined their sight lines during the conversation. “Sight line” here means the amount of time when they are looking at their interlocutor. In Fig.10, the vertical axis shows the proportion of time in the whole conversation for the subjects and the horizontal axis shows when the subject was looking at his interlocutor or the screen, for the patient and the student. In the case of showing photos, both the patient and the partner tended to look at their interlocutor. But in the case of showing videos, both of them tended to look at the screen.

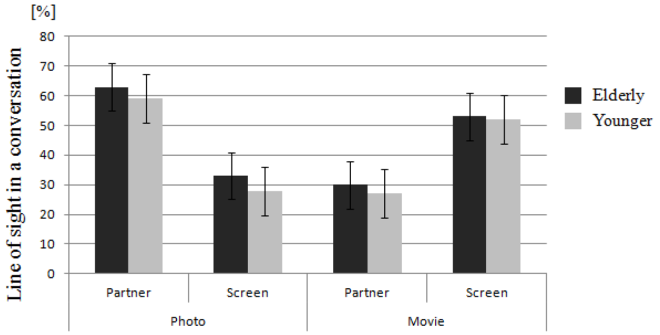


Fig. 10. Comparison of the line-of-sight during the speech time for the elderly and the young people

IV. Consideration

We found that there was a difference in the length of the communicative exchange for patients when viewing photos versus viewing videos. This must be caused by the fact that patients tend to watch videos more than their listening partners, which made interfered in the conversation. We expected in such a case, the communicative exchange on the part of the partner would become more frequent and longer, but it did not. They showed almost the same time in their communicative exchange in photo viewing and in video viewing. Still, we found showing photos made both of them talk longer.

We consider that one of the reasons why the patients talked longer when viewing photos is, through our watching the recorded video of the experiment, the patients gazed at the photo and talked about the photo shortly, and after that they talked for a long time about unrelated things. On the contrary, the patients watched the videos for a longer time, and talked less. But the student partner felt more interested in the video, thinking that they had succeeded in providing the topic which made them feel less burdened.

Thus we conclude that showing photos made both the patients and the partners talk at greater length. We suggest that the patients could talk more deeply on the topic by showing photos considering the former results of subjective evaluations, and that the student partner, as in the former results of the subjective evaluation, feel less burdened by showing videos, leading them to talk more intimately.

V. Challenges for the Future

We must examine in the future the mental burden of the partner in a given communicative exchange or in searching for topics. We anticipate constructing a support system which changes the content from photo to video properly adapted to the situation to reduce the volunteer’s mental burden.

As pointed out by the former study [4], when the students searched for photos or videos during the conversation, they felt that it was difficult to call a proper trigger word to mind. We suggest making the lists of each category like food, events, or play in order to remove the nuisance of searching for targets from an infinite set of data in YouTube or other pictures. Or at the point the list of the searched result is made, an easy previewing would help in choosing the one.

We intend to develop a touch panel system which would make searching for photos and/ or videos much easier.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15H01698

References

[1] A Overview Ministry of Health, Labour and Welfare, a 25-year Heisei version of Annual Report on Health and Welfare,Japan's population trends, the Ministry of Health, Labour and Welfare website Appendices (online) < http://www.mhlw.go.jp/wp/hakusyo/kousei/13-2/dl/01.pdf> pp.5

[2] Arlene J. Astell, Maggie P. Ellis, Lauren Bernardi, Norman Alm, Richard Dye, Gary Gowans, Jim Campbell, Using a touch screen computer to support relationships between people with dementia and caregivers;Interacting with Computers 22, pp.267-275(2010)

[3] Kuwahara Noriaki, Kuwahara Kazuhiro, Abe Shinji, Susami Kenji, Yasuda Kiyoshi; Video memories that utilize annotation of photo--Application and evaluation to persons with dementia - - making support; Artificial Intelligence Journal, Vol.20, No.6, pp.396-405 (2005)

[4] Tuji Airi, Kuwahara Noriaki, Morimoto Kazunari; Implementation of interactive reminiscence photo sharing system for elderly people by using web services; Human Interface Society, (2010)

[5] Takai Shota; Topic suggestions propulsion system of conversation during the first meeting in accordance with the TPO; Heisei 21 year master's thesis, (2010)

[6] Iwamoto Miyuki, Kuwahara Noriaki, Morimoto Kazunari; Comparison between the Burden of the Conversation by Using Photographic Image and that by Using Motion Video; Human Interface Society 2012, pp.579-584 (2012)

Medical Image De-Noising Schemes using Wavelet Transform with Fixed form Thresholding

Nadir Mustafa[1]
[1]School of Computer Science &Technology,
UESTC, Chengdu, 611731, China

Jiang Ping Li[2]
[2]School of Computer Science & Technology,
UESTC, Chengdu, 611731, China

Saeed Ahmed Khan[3]
[3]Department of Electrical Engineering, Sukkur Institute of
Business administration Sindh, Pakistan

Mohaned Giess[4]
[4]School of Communication & Information Engineering,
UESTC, Chengdu, 611731, China

***Abstract*—Medical Imaging is currently a hot area of bio-medical engineers, researchers and medical doctors as it is extensively used in diagnosing of human health and by health care institutes. The imaging equipment is the device, which is used for better image processing and highlighting the important features. These images are affected by random noise during acquisition, analyzing and transmission process. This condition results in the blurry image visible in low contrast. The Image De-noising System (IDs) is used as a tool for removing image noise and preserving important data. Image de-noising is one of the most interesting research areas among researchers of technology-giants and academic institutions. For Criminal Identification Systems (CIS) & Magnetic Resonance Imaging (MRI), IDs is more beneficial in the field of medical imaging. This paper proposes an algorithm for de-noising medical images using different types of wavelet transform, such as Haar, Daubechies, Symlets and Bi-orthogonal. In this paper noise image quality has been evaluated using filter assessment parameters like Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE) and Variance, It has been observed to form the numerical results that, the presentation of proposed algorithm reduced the mean square error and achieved best value of peak signal to noise ratio (PSNR). In this paper, the wavelet based de-noising algorithm has been investigated on medical images along with threshold.**

Keywords*—Image De-noising System; GUI De-noised image; Code De-noised image; Wavelet transform; Soft and Hard Threshold

I. Introduction

Recently most of human-assistedcomputer applications rely on the use of digital image processing techniques, such as magnetic resonance imaging (MRI), criminal identification systems (CIS), agricultural and biological research (ABR). The term image de-noising is the best tool used in these applications, where it aims at remove the noise and retain important image features as much as possible. The use of medical imaging (MRI) in diagnosis has been greatly accepted for its non-sensitive features, low cost, the ability of constructing real-time image with improved property[1], [2].During image acquisition and transmission, it has been usually observed that random noise always occurs at another end. So this noise causes problems such as a blurred vision of images, which reduce the visuality of low-contrast articles. Therefore, it is not easy for the medical doctors to examine the abnormalities in human in the invisible image. The process of removing noise is necessary in most medical imaging equipments for the purpose of enhancing miniatures that may be concealed in the data [3][4].

II. Wavelet Transform

This wavelet transform is alike to Windowed Fourier Transform (WFT), but themerit function is totally different. The main difference between the Window Fourier Transform and wavelet lies in the signal analysis; The WFT breaks down the signal into cosines and sines and, namely, the functions are restrained in Fourier space. On the contrary, functions that are utilized in the wavelet transform are confined in the real space and the Fourier space. Commonly, the Continuous Wavelet Transform (CWT) is containing different parameters which are derived from Fourier analysis transform and mother wavelet transform. The equation (1) describes the parameter $\gamma(s,\tau)$ is a wavelet coefficient with scale s and timeτ, and the function $f(t)$ is define as the time series wherethe certain function is $\psi_{s,\tau}^{*}$ defines a complex conjugate of wavelet with scale and timess,τ.[5][6].

$$\gamma(s,\tau) = \int f(t)\psi_{s,\tau}^{*}(t)dt \qquad (1)$$

Wavelets have been considered recently as a strong tool for de-noising image. The individual wavelet makes an image into a group of coefficients that compose a multi-scale model of the image. The distinct wavelet transform of signal expressed as x(n) is calculated by making it go through a low pass filter with impulse response g(n) as long as given an approximation coefficient. The signal is breaks down concurrently by the use of a high pass filter h(n), while gives details coefficients. These filters are named asQuadratic Mirror Filters. Because thehalf of frequencies of the signal is taken out, the sample of the filter outputs are reduced by equation (2)&(3).

$$Y_{low}[k] = \sum_{n} x[n].g[2k-n] \qquad (2)$$

$$Y_{high}[k] = \sum_{n} x[n].g[2k-n] \qquad (3)$$

Image is a 2-dimentional signal, and we use x (N, M) to represent it. Firstly each row is filtrated and then down-sampled to get two images represented by (N, M/2), secondly every column is filtrated and down-sampled toget four sub bands named as HH, HL, LH and LL Therefore, in case of two dimensions, one 2-D scaling function and three 2-D wavelet functionsare generated.

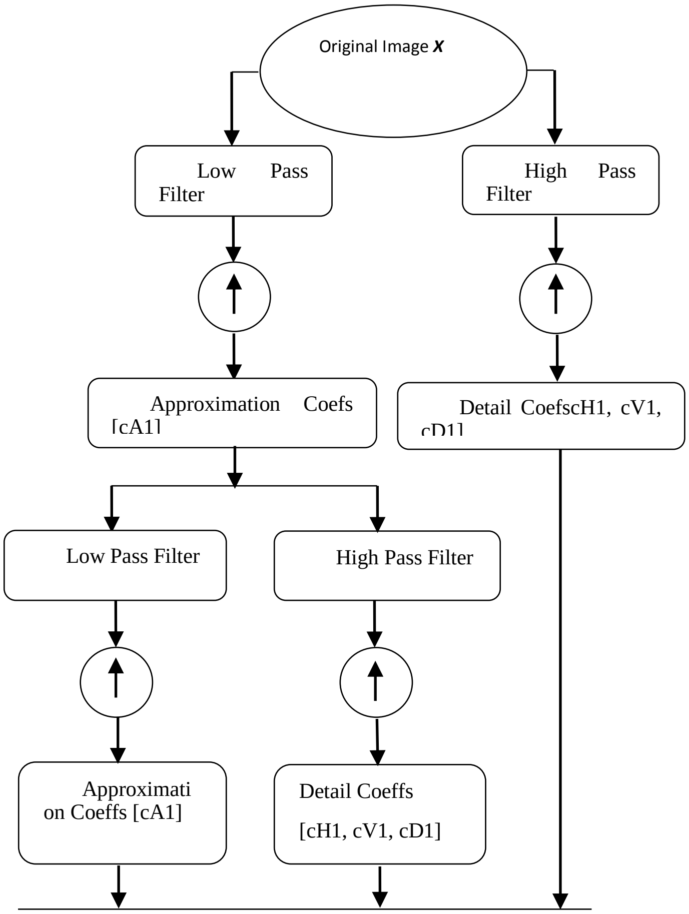


Fig. 1. The 2D discrete wavelet decomposition

The variable CA1, CD1, CH1 and CV1 stand for approximation coefficient, diagonal detail coefficient, horizontal detail coefficient and vertical detail coefficient.

At level two approximation sub-band LL is then decomposed into four components, the performance can be progressed all the same for anotherthree levels. LL has strength concentration for low pass and HH sub-band for high-frequency constituents. Rebuilding can be performed by IDWT (Inverse Discrete Wavelet Transform) to obtain the de-noised image [7] [8].

The process of 2D discrete wavelet decomposition has been depicted in figure.1, which describes the main steps for de-noising. The process starts from image decomposition, up-sampling and down-sampling until the reconstruction of four sub band coefficients are obtained for original image [9][10].

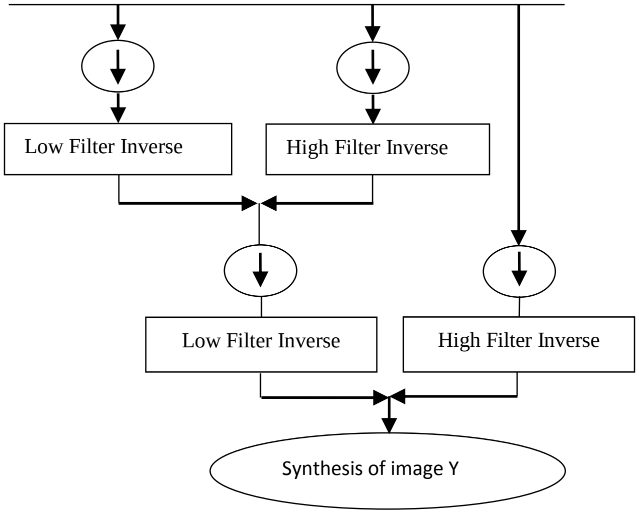


Fig. 2. The 2D discrete wavelet reconstruction

Here, figure.2 illustrates; brain image reconstruction from three-level decomposition. We can see the wavelet decomposition process can be seen byconsecutive approximations being decomposed successfully. In figure three the original medical image shows decomposition into many elements with lower-resolution.

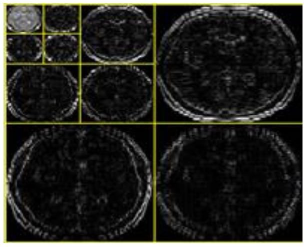


Fig. 3. Wavelet three-level decomposition of brain image

III. Proposed Algorthim for De-Noising

The purpose of this paper is the de-noising of medical image of the brain usingdifferent types of wavelets, such as Haar, db10, sym3 and bior3.7 wavelet. Our contribution in this paper is that good results are obtained when applying fixed form threshold in terms of soft and hard threshold algorithm. To evaluate the proposed algorithm, several parameters are used such as Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE) and Variance. Numerical results show the validity of proposed algorithm. The mean square error is reduced, while a peak signal to noise ratio (PSNR) is achieved.

A. Image De-noising Algorithm

There are three steps of de-noising proceduredescribed as follows:

Wavelet decomposition level Pick a level (level-3). Calculate the wavelet decomposition of the noisy image at level 3. The wavelet produces all the coefficients, from the wavelet analysis process.

Threshold detail coefficients:a threshold is chosen for level3 and softthresholdingisapplied to the detail coefficients. If the wavelet coefficients are larger than the threshold value, those coefficients are leftunaltered. If they are small than threshold, they are restrained.

Reconstruct wavelet coefficients based on level 3 of wavelet transform. Then, transformdetailed coefficients from level 3 to level 1.

B. ThresholdingParameter

In this part parameters are formulated andused for de-noising.

1) Noise variance

Apply a fixed form thresholding algorithm to the wavelet coefficients. In fixed form, the noise variance is calculated using the median of absolute deviation of the transform coefficient of all three levels; the (MAD) is given by equation (4).

$$\sigma_n^2 = \frac{median(abs(x_{ij}))^2}{0.6745} \qquad (4)$$

2) Threshold Parameter

The threshold (T_h) is a threshold parameter applied to wavelet coefficients of a noisy image. Where M is number of pixels in theimage,and S is the noise variance and the threshold is given by equation (5).

$$Th = \sigma\sqrt{2\log M} \qquad (5)$$

Hard thresholding is a keep or kill the wavelet coefficients compared with threshold parameter. The threshold is deducted from any coefficient that is larger than the threshold. This process makes the time series move toward zero.

C. Evaluation Parameters

In this partevaluation parameters are discussed.

1) Mean Square Error (MSE)

The MSE estimate the quality alteration between the GUI de-noised image (X)and code demised image (Y), the average of the squared image is given in equation (6).

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[Y(i,j)-X(i,j)]^2 \qquad (6)$$

2) Peak Signal to Noise Ratio (PSNR)

The PSNR represents the size of the error in relation to the peak value of the signal rather than the size of the error in relation to the average squared value of the signal. It is computed with the size of the error in relation to the average squared value of the signal. PSNR is greater for a better-transformed image and smaller for a poorly transformed image. PNSR calculates image fidelity, i.e., intimately the transformed image looks like the initial image, the PSNR exhibited in equation (7).

$$PSNR = 10\log_{10}\left[\frac{S^2}{mse}\right] \qquad (7)$$

IV. Research Methodology

The experiments in this paper have been conducted on two medical images; of Brain with different size. The first image is a brain medical image with size $[204x204]$, the second image is a brain medical image with size $[150x150]$; Different types of wavelet transform have been applied respectively (haar, db10, sym3, and bior3.7) for these two images to generate de-noised image. After applying wavelet, (CA) approximation and (CD) details coefficient at three levels of decomposition process have been generated. These coefficients represented in vector [C,S] such as [CD1, CD2, CA3, CD3]. After each level consists of horizontal, vertical and diagonal coefficients, de nosing image is achieved. As there are many threshold levels but in this paper, fixed form soft threshold for three levels of decomposition process have been selected because it will give best threshold value. Here un-scaled white noise is added to the original image to generate a de-noising image in GUI (Graphical User Interface). At the first stage, the original image is compared with the GUI de-noising image. for the same scheme, MATLAB codesare written to compare the original image using hard threshold with the image de-noising generated code. At the later stage the GUI de-noising image is compared with the image de-noising generated code along with MSE and PSNR parameter.

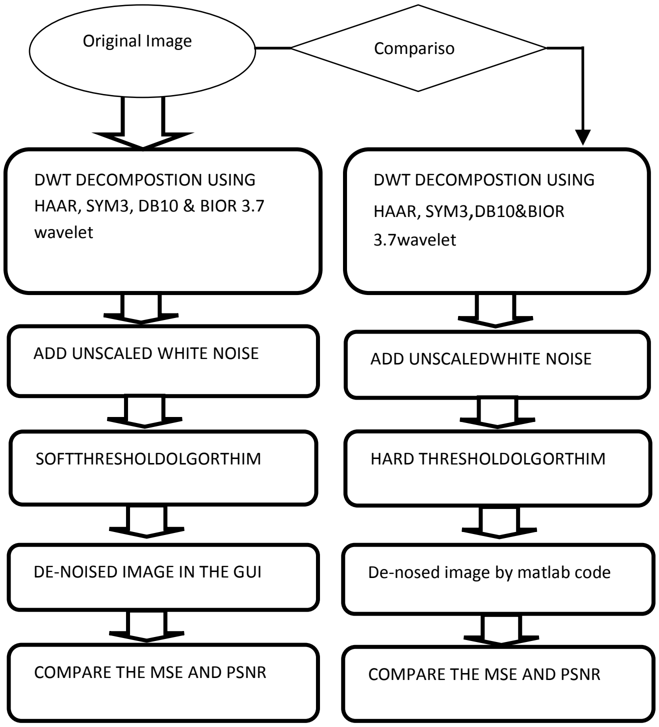

Original Image
Compariso
DWT DECOMPOSTION USING HAAR, SYM3, DB10 & BIOR 3.7 wavelet
DWT DECOMPOSTION USING HAAR, SYM3,DB10&BIOR 3.7wavelet
ADD UNSCALED WHITE NOISE
ADD UNSCALEDWHITE NOISE
SOFTTHRESHOLDOLGORTHIM
HARD THRESHOLDOLGORTHIM
DE-NOISED IMAGE IN THE GUI
De-nosed image by matlab code
COMPARE THE MSE AND PSNR
COMPARE THE MSE AND PSNR

Fig. 4. De Nosing Comparison Algorithm Model

V. Results and Discussion

Here Fig.5 illustrates the initialmedical image of brain image. Fig 6&7 depicts the de-noising images generated in GUI and MATLAB Code.

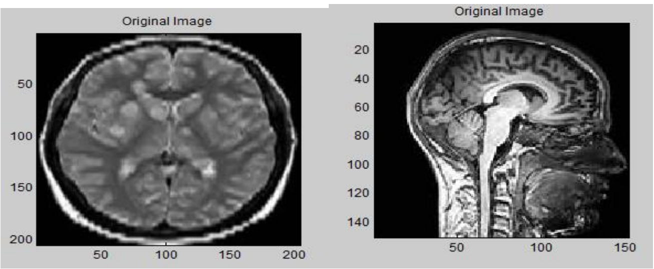


Fig. 5. The original of two brain images

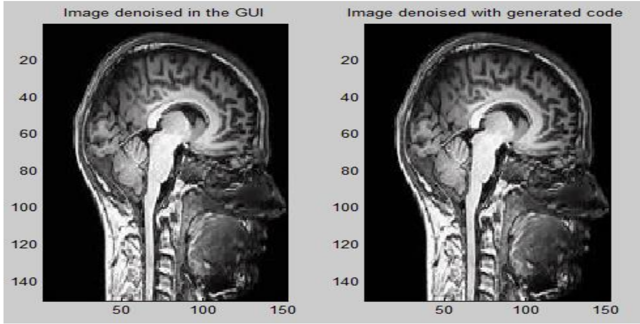


Fig. 6. The De-noised of first brain image

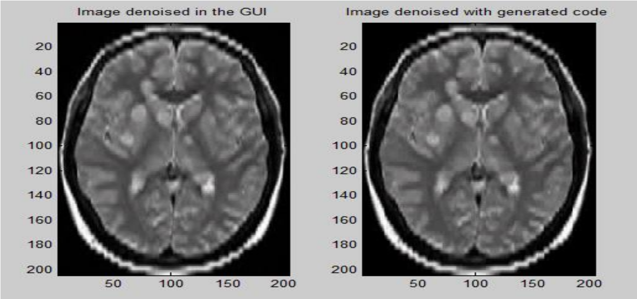


Fig. 7. The De-noised second brain image

The Performance evaluation of de-noising image can be observed in the tables 1 & 2. From the relationship of the peak signal and mean squareto the noise ratio, here it can be observed from Table 1& 2 that the MSE measurements in the GUI generated de-noising image are smaller than the measurements in the code generated de-noising image in all wavelet families. It reveals that the mean square error of the initialimage in GUI is less than the hard threshold generated code of the original image. That is because of the image size.

TABLE I. PERFORMANCE EVALUATION OF GUI METHOD FOR DIFFERENT THRESHOLDING IN TERM OF MSE, PSNR FOR TWO DIFFERENT BRAIN IMAGES

Various image	Wavelet package	Soft Threshold Method		Hard Threshold Method	
		MSE	PSNR	MSE	PSNR
Brain image with size (150×150)	haar	0.8436	48.2026	3.4461	41.7699
	db10	0.8697	48.2200	3.2437	41.6534
	sym3	0.7360	48.2309	3.2600	41.8367
	bior3.7	0.6589	48.2433	2.6374	41.9085
Brain image with size (200×200)	haar	0.8280	48.1395	4.3343	38.9221
	db10	0.8595	48.1526	5.5580	39.9631
	sym3	0.7062	48.1809	4.8423	40.2849
	bior3.7	0.6612	48.2085	3.4412	42.2548

TABLE II. PERFORMANCE EVALUATION OF MATLAB CODE METHOD FOR DIFFERENTTHRESHOLDING IN TERM OF MSE, PSNR FOR TWO DIFFERENT BRAIN IMAGES

Various Images	Wavelet package	Soft Threshold Method		Hard Threshold Method	
		MSE	PSNR	MSE	PSNR
Brain image with size (150×150)	haar	0.9080	48.1395	7.3343	38.9221
	db10	0.9395	48.3085	4.5580	39.9631
	sym3	0.8962	48.3709	5.8423	42.2849
	bior3.7	0.7212	48.6485	2.4412	44.2548
Brain image with size (200×200)	Haar	0.9312	47.2095	9.4413	38.3264
	db10	0.9122	47.2526	7.2052	40.5231
	sym3	0.9482	47.3929	8.2311	41.4223
	bior3.7	0.7551	47.5425	3.1220	43.1253

Fig. 8 & 9 illustrates the relationship of MSE & PSNR of four wavelet families for brain de-noising medical image. Here it can be observed that bior3.7 wavelet has better results than the other wavelet families used in this paper for image de-noising.

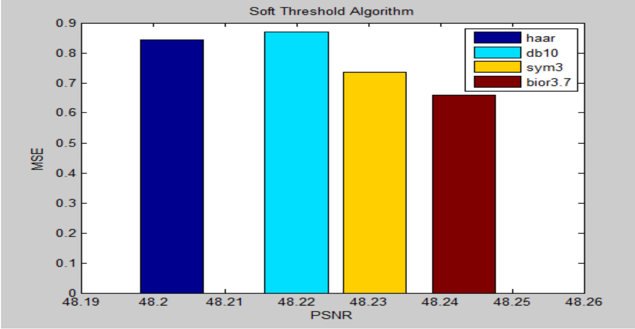


Fig. 8. The histogram of GUI method using Soft Threshold Algorithm for Brain de-noising image

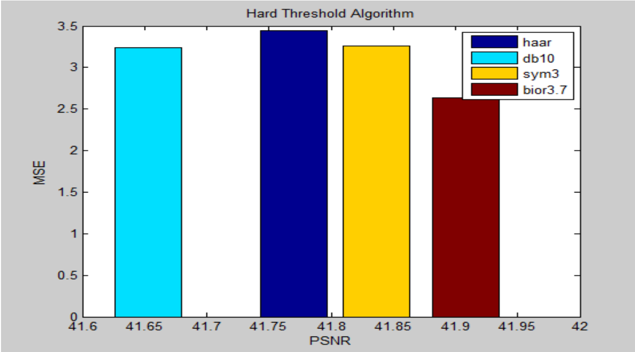


Fig. 9. The histogram of GUI method using Hard Threshold Algorithm for Brain de-noising image

Fig. 10 & 11 illustrates the relationship of MSE & PSNR of four wavelet families for the brain de-noising medical image. Here it can be observed that bior3.7 wavelet has better results than the other wavelet families used in this paper for image de-noising.

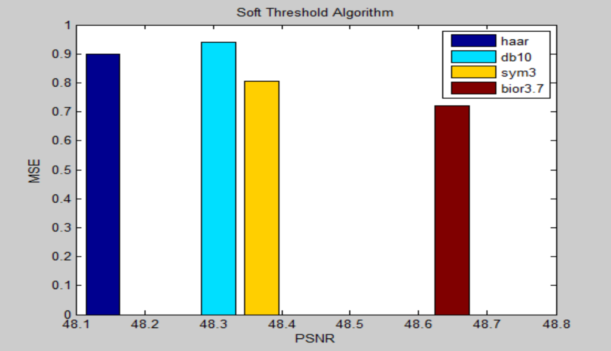


Fig. 10. The histogram of MATLAB code method using Hard Threshold Algorithm for brain de-noising image

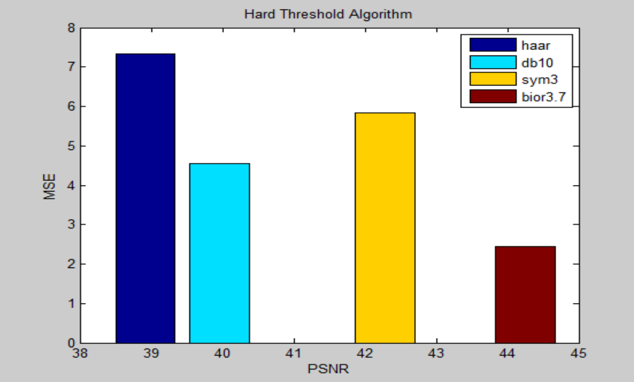


Fig. 11. MATLAB code method using Hard Threshold Algorithm for brainde-noising image

Fig. 12& 13 illustrates the relationship of MSE & PSNR of four wavelet families for the brain de-noising medical image. Here it can be observed that bior3.7 wavelet has better results than the other wavelet families used in this paper for image de-noising.

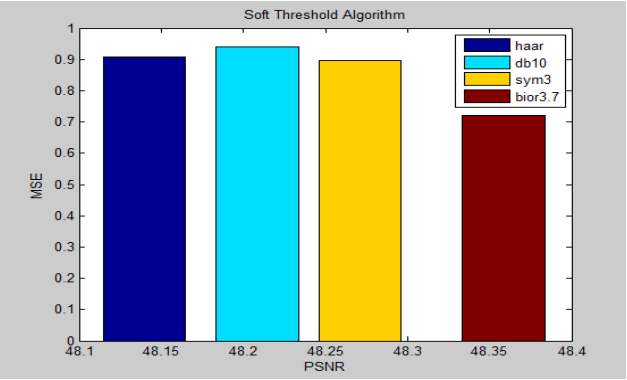


Fig. 12. The histogram of GUI method using Soft Threshold Algorithm for brain de-noising image

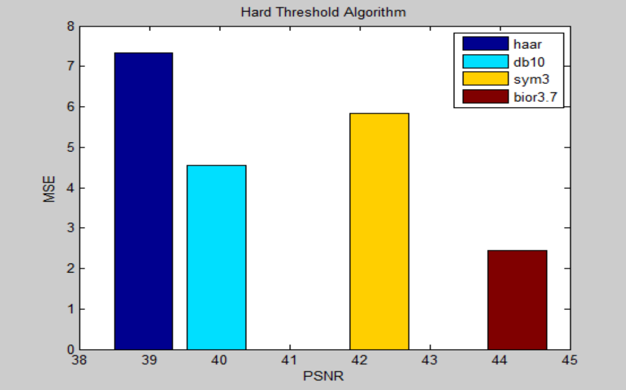


Fig. 13. The histogram of GUI method using Hard Threshold Algorithm for brain de-noising image

Fig. 14& 15 illustrates the relationship of MSE & PSNR of four wavelet families for the de-noising medical image. Here it can be observed that bior3.7 wavelet has better results than the other wavelet families used in this paper for image de-noising.

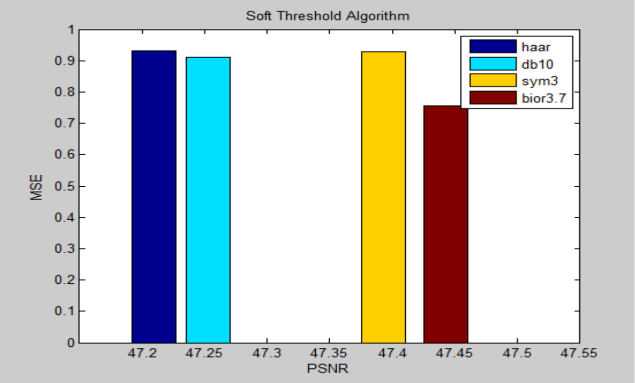


Fig. 14. The histogram of MATLAB code method using Hard Threshold Algorithm for Brain de-noising image

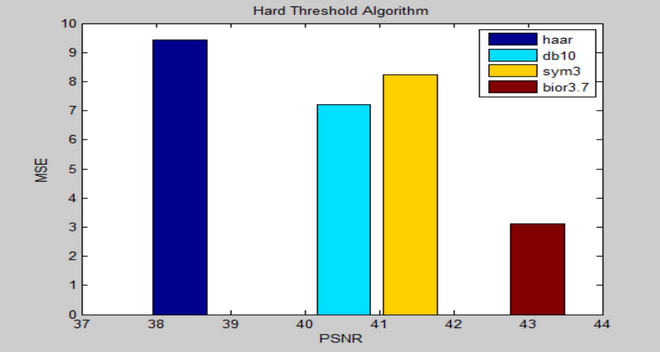


Fig. 15. The histogram of MATLAB code method using Hard Threshold Algorithm for Brain de-noising image

VI. Conclusions

In all images, noise is the main problem, and one has to nip this problem in the bud for better results.

De-noising is very crucial especially in medical science. In this paper, removing the un-scaled white noise added to original medical images has been presented. The new algorithm has been proposed for De-noising brain medical images. The proposed new algorithm on the basis of the wavelet transform is observed to be a more competent method in image de-noising especially forremoving un-scaled white noise. Qualitative and quantitative analysis results reveal that the proposed algorithm reduces the mean square error (MSE) of different images with different sizes using different wavelet families for hard and soft threshold. Experimentsrepresent that, the bi-orthogonal wavelet is a more efficient method than other wavelet families discussed in this paper, such as Haar, Daubechies, and Symlets because it gave better results with mean square error (MSE) in soft and hard threshold. Efficient de-noising values in a soft threshold algorithm are generated in GUI. Because of difference in image sizes, hard threshold algorithm generated code values were observed as larger in case of the brain medical image. Results reveal that bi-orthogonal wavelet shows the best results with parameter MSE and PSNR. At the later stages one can work precisely on MSE and PSNR measurements for both soft and hard thresholds for getting de-noised medical images.

ACKNOWLEDGEMENT

This paper was supported by the National Natural Science Foundation of China (Grant No: 61370073), and the National High Technology Research and Development Program of China (Grant No: 2007AA01z423).

REFERENCES

[1] A. Bultheel: Bull. Belg. Math. Soc.: (1995) 2

[2] S.Satheesh, Dr.KVSVR Prasad, “Medical image de-noising using adaptive threshold based on contourlet transform” , An International Journal (ACIJ), Vol.2, No.2, March 2011.

[3] S. G. Chang, B. Yu, M. Vetterli: IEEE Trans. Image Processing, (2000) 9 p. 1532

[4] S. G. Chang, B. Yu, M. Vetterli: IEEE Trans. Image Processing, (2000) 9 p. 1522

[5] Sivakumar .R and Nedumaran .D, “ Implementation of Wavelet Filters for Speckle Noise Reduction in Ultrasound Medical Images: A Comparative Study” International Conference on Signals, Systems and Communication, December 21-23, 2009.

[6] AsmaaAbassAjwad , “ Noise Reduction of Ultrasound Image Using Wiener filtering and Haar Wavelet Transform Techniques” Diyala Journal of Medicine, Vol. 2, Issue 1, April 2012.

[7] Dhruv Kumar, MaitreyeeDutta, ParveenLehane, “ A Comparative Analysis of Different Wavelets for Enhancing Medical Ultrasound Images”, International Journal

[8] Eslami, R. and Radha. H., “Translation-invariant contourlet transform and its application to image denoising,” IEEE Transactions on Image Processing, Vol. 15, No.11, pp. 3362–3374, 2006

[9] M. N. Do and M. Vetterli, “Pyramidal directional filter banks and curvelets,” Proc. IEEE Int. Conf. on Image Proc., vol. 3, 2001, pp. 158-161.

[10] M. N. Do and M. Vetterli, “Contourlets: a Directional Multiresolution image representation,” Proceedings of 2002 IEEE International Conference on Image Processing, vol. 1, 2002, pp 357-360.

Automation and Validation of Annotation for Hindi Anaphora Resolution

Pardeep Singh
Computer Science and Engineering
National Institute of Technology
Hamirpur, INDIA

Kamlesh Dutta
Computer Science and Engineering
National Institute of Technology
Hamirpur, INDIA

***Abstract*—The process of labelling any language genre by which one can extract useful information is called annotation. This provides syntactic information about a word or a word phrase. In this paper, an effort has been made to provide the algorithm for semiautomatic annotation for Hindi text to cater anaphora resolution only. The study was conducted on twelve files of Ranchi Express available in EMILLE corpus. The corpus is originally tagged for demonstrative pronouns. The detection of the pronouns is supported by the incorporation of seven tags. However the semantic interpretation of the demonstrative pronoun is not supported in the original corpus. In this paper an effort has been made to automate the process of tagging as well as the handling of semantic information through addition tags. It was conducted on 1485 demonstrative pronouns. The average accuracy of precision, recall and F measure is 74, 71 and 72 respectively.**

Keywords—Annotation; natural language processing; demonstrative pronoun; semantic category; indirect anaphora; semiautomatic annotation

I. Introduction

Natural language processing has attracted the researchers' volition to enhance the natural language resources during the last few decades. A number of applications of natural language processing need the syntactic meanings of words or word phrases. These meanings are used for different applications like, part of speech tagging, information retrieval, text summarizations, question answering, anaphora resolution, etc. So, the annotation will play a pivotal role in these NLP applications. In this study, the systematic discussion has been held of the labelling process of demonstrative pronoun in the context of anaphora resolution.

Anaphora is a process of finding the referring expression in the discourse. Wrong correlation of referring expression in the genre affects all applications of NLP.

Example 1:

"***They*** don't understand why it seems like bad behavior on Wall Street is rewarded, but hard work on Main Street isn't, or why Washington has been unable or unwilling to solve any of our problems", (Obama 2010, http://www.diva-portal.org/smash/get/diva2:531167/fulltext01.pdf, accessed on 18th Aug, 2015).

In example 1, Obama referred to the people in a negative way by implying that they might be inferior, since Obama assumed that they do not understand how the economic crisis was solved. Though the Obama wants to refer the Congress rather people.

Example 2:

"Now, our friends down in Tampa at the Republican Convention were more than happy to talk about everything ***they*** think is wrong with America. But ***they*** didn't have much to say about how ***they'd*** make it right. ***They*** want your vote, but ***they*** don't want you to know their plan. And that's because all ***they*** have to offer is the same prescriptions ***they've*** had for the last 30 years" (Obama, Sept 6, 2012, http://www.presidency.ucsb .edu/ws/index.php?pid=101968, accessed on 18th Aug, 2015.)

In example 2, '***they'*** refer to a specific group, which does not belong to Democrats and this demonstrative pronoun creates serious problems in anaphoric context. Moreover the interpretation of machine has been always error prone.

So, anaphora resolution itself is a significant problem. To address this problem annotation of any corpus is crucial, while formulation, evaluation and optimization of any algorithms in NLP, particularly automation of anaphora resolution. Annotation becomes the prerequisite condition for anaphora and other applications for better accuracy.

II. Background of Annotation

A number of attempts have been made to retrieve the information from the text by a number of means; one of them is annotation. There is no standard annotation scheme which can fulfill all the requirements. The different labeling schemes have been adopted to address the different problems. In this regard the most commonly used practices are phrase structure, dependency, HPSG (Head-driven Phrase Structure Grammar) and Hybrid (Phrase structure and Dependency, both). Penn Treebank [1] is the most used and adapted annotation scheme; firstly for English and then in other languages. A number of languages parsed according to Pen Treebank are like, Arabic, Bulgarian, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, etc.

III. Motivation

Annotated corpora promise to be valuable for researchers as diverse as the automatic construction of statistical models. Written or spoken language provides the raw data to investigate, evaluation and comparison of different linguistic tools/ models. Combining raw language data with linguistic information, offers a promising basis for the development of new efficient and robust NLP methods. Real world texts annotated with different strata of linguistic information can be used for grammar induction. Annotating the corpora manually takes rigorous effort and competency too. It is better to draw some conclusion/ rules which lead to fully or partially automate this process.

Skill and competency levels of human beings always impose the restriction on the accuracy of annotation. Human interpretation of discourse and its constituents may vary. Understanding may become subjective in the context and may lead to incorrect annotation. Though, manually tagged data are the most preferred practice, may be due to the ultimate benchmark for accuracy of any NLP task, i.e., human interpretation, the need for having automation is necessitated by the availability of voluminous digital data these days and the limitation imposed by the human capacity.

IV. LITERATURE SURVEY

A number of attempts have been made to label the text or dialogue. Broadly, these are classified under four categories. First dependency structure, second phrase structure, third HPSG (Head-driven Phrase Structure Grammar) and fourth is hybrid labeling. Technically, hybrid type of annotation is to serve two purposes; phrase structure as well as dependency. These annotation schemes are being used to enrich language resources. Multiple corpuses for same language is an added advantage for researchers to test and validate their tool/ technique for the same. It helps to create a more generic solution for those languages.

A. Manually Annotated Corpora

Fully tagged corpus is a necessity of any language to groom its automation tools. A number of corpora are available in English and European languages. The Prague Dependency Tree Bank (PDT) annotated up to three levels; morphological, syntactic annotation and third level were linguistic level [1]. Susanne Corpus as treebanks exist for English [2], the Lancaster Parsed Corpus [3] and another corpus for English, the International Corpus of English [4], the Prague Dependency Treebank for Czech [5]. Treebank projects for other languages made in the recent times, e.g., for French which is tagged for morphosyntax, lemmas, compounds, lexical clusters and phrase boundaries [6], Italian corpora annotated with grammatical relations and syntactic representation of sentence [7], syntactically parsed for Spanish [8], lexically annotated speech corpora of Turkish was an attempt by marking derivation boundaries by [9], and a dependency structured Russian corpora which was lemmatized, morphologically and syntactically tagged [10]. The annotation of the German TIGER Treebank [11] is done in a manner so that it can easily be exported to XML. They consider the verb-sub categorization, coordination, appositions and parentheses as well as proper nouns. TIMEBANK is richly annotated to indicate events, times, and temporal relations [12].

B. Annotated corpora for Coreference and Anaphora

There are a few corpora which were annotated, especially for anaphora resolution or co-reference resolution. Lancaster Anaphoric Treebank [13] of Associated Press with 100000 words and annotated according to UCREL annotation scheme. This was the joint venture of UCREL and IBM. MUC-6 and MUC-7 annotate co-referential link of 65000 words. Similar to MUC scheme a tool ClinKa was developed to annotate English genre at University of Wolverhampton [14]. Another corpus developed by the members of University of Stendahls Grenoble and Xerox Research Center Europe [15] by creating an anaphoric and cataphoric link. It addresses the zero noun anaphora, adverbial anaphora, indefinite pronoun, demonstrative pronoun, third person personal pronoun, and personal pronoun. French corpus [16], annotate anaphoric links in MUC. Few multilingual corpora are available like English-Romanian corpus [17], technical manual of English- French for co-referential link at the University of Wolverhampton.

C. Automatic or Semiautomatic Annotated Corpora

An attempt for corpus annotation for labelling semantic and syntactic meaning of word and word phrases for coverage of deep parser to generate syntactic structure, semantic representation and discourse information for dialogue by means of semiautomatic technique [18]. The EPAC was a speech corpus, it consists of a set of 100 hours of conversational speech manually transcribed [19]. This spoken corpus automatically annotated by automatic segmentation, transcription, POS tagging and other tools. The Diachronic German Corpus [20] was automatically annotated by a suite of NLP tools. These tools are integrated into WebLicht and CLARIN-D. WebLicht Service Oriented Architecture is used as an integrated environment. A corpus is trained with automatic system for semantic [21]. It advocates coreference, quantification, and defined a set of semantic rules for many other higher-order phenomena, which was left out by Penn Treebank.

D. Annotation for Hindi

Botley & Mc Energy [22] proposed an annotation scheme for English to resolve anaphora. Later this scheme modified by Botley [23] again for the same purpose and same language with emphasis on indirect anaphora. They considered the recoverability of antecedent, direction of reference, phoric type, syntactic function, antecedent type to annotate three genres. These corpora are the American Printing House for the Blind (APHB) Corpus, the Associated Press (AP) Corpus, and the Hansard Corpus [23]. Recoverability refers to the relation between referring expression and its corresponding antecedent in context of demonstrative pronoun and this feature based on Halliday and Hassan [24]. Feature “Phoric type” is distinction between substitution and reference [24]. Their tag set adapted to annotate Hindi by Sinha [25], and later [26] added a few more tags. Reference [27] used some semantic information for indirect anaphora categorization. The author has identified a few semantic categories to classify indirect anaphora in Hindi.

V. METHODOLOGY

A. Tag set used

The demonstrative pronouns are understood in terms of an unordered paradigmatic set of five distinctive features [22] . Syntactic Function and Antecedent Type, two other features, which are proposed by [23]. Last three features in table 1 facilitate to identify indirect anaphora [27], [26], [25]. Recoverability of antecedent helps to identify the same.

TABLE I. FEATURE USED FOR ANNOTATION

No. of Feature	Feature	Value1	Value2	Value3	Value4	Value5
1.	Distance Marking	P (proximal)	D (Distal)	None	None	None
2.	Nature Of deixis	P (Pronoun)	D (Demonstrative)	Z (Zero)	None	None
3.	Recoverability of Antecedent	D (Directly Recoverable)	I (Indirectly Recoverable)	N (Non-recoverable)	0 (not applicable, e.g. exophora)	None
4.	Direction of reference	A (anaphoric)	C (cataphoric)	0(not applicable, Exophoric or deictic)	None	None
5.	Phoric Type	R (Referential)	0 Not Applicable	None	None	None
6.	Syntactic Function	M (Noun Modifier)	H (Noun Head)	0 (Not Applicable)	None	None
7.	Antecedent Type	N (nominal)	P (propositional/ Factual)	C (Clausal)	J (Adjectival)	O (None)
8.	Pronoun pattern	Pronoun and subsequent construct in the sentence				
9.	Case marker/ Connective	Case marking or connective following the pronoun				
10.	Semantic/ category	Semantic categories				

B. *Corpus selection*

This study has been conducted on EMILEE corpus. The Hindi written corpus contains a total of approximately 12,390,000 words in Unicode. It is pre annotated corpora for demonstrative pronouns. One component of this corpus is based upon Ranchi Express news items. In pre annotated corpus there are seven features. We have modified this corpus with an annotation scheme according to [27]. Being in Unicode is an added advantage of using EMILEE corpus.

Example 1 is manually annotated Ranchi Express news from EMILEE corpus according to Botley's annotation. In this annotation only seven tags are there and their respective value of उन्होंने (unhone) pronoun is DPDARHN.

Example 3:

<body>
<p>किसी मंत्री को
हटाने का सवाल नहीं : मरांडी</p>
<p>रांची: मुख्यमंत्री बाबूलाल मरांडी ने आज
विधानसभा में कहा कि पलामू में एक लड़की के अपहरण की घटना के क्रम में झारखंड मंत्रिमंडल से किसी सदस्य को हटाने का सवाल ही पैदा नहीं होता। <w tag = " DPDARHN"> उन्होंने</w> कहा कि <w tag= " P D DARMN">यह</w> मामला कई दिनों से चर्चा में है लेकिन, घटना अपहरण की है अथवा लड़का और लड़की स्वेच्छा से गए हैं <w tag="PDDARHC">यह</w> जांच का विषय है।

We have considered the data from EMILLE corpus. We picked one segment of corpus which is based on the news items from Ranchi express. In this study, we analysed twelve files of plain text.

Example 4: Manually annotated Ranchi Express news from EMILEE corpus according to table 1.

<body>
<p>किसी मंत्री को
हटाने का सवाल नहीं : मरांडी</p>
<p>रांची : मुख्यमंत्री बाबूलाल मरांडी ने आज विधानसभा में कहा कि पलामू में एक लड़की के अपहरण की घटना के क्रम में झारखंड मंत्रिमंडल से किसी सदस्य को हटाने का सवाल ही पैदा नहीं होता। <w tag= "D, P, D, A,R,H,N,उन्होंने,यह,_,_,_">उन्होंने</w> कहा कि <w tag="P,D,D,A,R,M,N,यह, यह,_,_,_"> यह</w> मामला कई दिनों से चर्चा में है लेकिन, घटना अपहरण की है अथवा लड़का और लड़की स्वेच्छा से गए हैं <w tag="P,D,D,A,R,H,C,यह,_,_,_">यह</w> जांच का विषय है।

In example 4 additional three tags have been attached उन्होंने,यह, ,_,_,_". First is pronoun pattern, second case marker and third semantic category. In this example null is denoted by '_'.

C. *Algorithm*

Though, some researchers advocated syntactic features of languages for annotation. Reference [27] suggested some specific pattern of pronoun and other words which is categorized in pattern. Secondly, case marker elaborates the pronoun significance and binding with its antecedent. This algorithm annotate only last three tags discussed in example 4.

```
Step1: Input the set of case marker.
Step2: Input pattern for pronouns
Step3: WHILE(file in the lists of files) REPEAT
        I)   Find "<w tag, and corresponding ">"
           a.  Extract the feature list call it tag
        II)  Split the tag list into list of individual
             features
        III) Generate_Case_Marker_and_Pronoun_
             Pattern()
           a.  Define window size for pattern
```

```
            b. Within window size, find the
               case_Marker and pattern_follow_Pronoun
            c. Extract case_Marker and Pattern_follow_
               Pronoun
        IV) Classification()
            a. Extract semantic category, which is tag
               [10], Pronoun, which is tag [8], pattern
               following pronoun which is stored in
               string tag [].
            b. Sequential search these in the rule based
               classifier.
            c. Output the CLASS; return CLASS.
    END WHILE
Step4: Print all the text along with modified tag to files.

Step5: Stop
```

EMILEE corpus is annotated with seven tags. Last three tags (Pronoun pattern, Case marker/ Connective, Semantic category) have to be annotated either manually or automatically. Reference [27] draw some rules for classification of indirect anaphora in demonstrative pronoun.

As a first step, the eight case markers of Hindi given as input then pronoun pattern. Third step is to read all files to be annotated. It extracts all tags (feature list) which is already annotated in EMILLE corpus. Then it generates the case marker and a pronoun pattern by defining window size. We took window size five. In this window, algorithm will search specific pronoun pattern and case marker according to [27]. Pronoun was matched with the given list of pronoun and proceeding word; which indicates recoverability of antecedent on the basis of semantics. We considered the first hit for pronoun pattern and case marker. Then extract the case marker and pronoun pattern. In fourth step syntactic category, pronoun and pattern following pronoun are stored as the element of a string tag (i.e. tag [8], tag [9] and tag [10] respectively). Then apply the rules given by [27]. In '*b*' part of '*fourth'* step of above algorithm (i.e. Sequential search these in the rule based classifier) has been adopted from [27]. The output is stored in one class. This is the required output with file for all ten tags.

D. *Case marker used*

Hindi consists of nine case markers. First eight are in use and the last one "hey" (हे) is obsolete, or less in use practically. These case markers specify the binding with anaphor and antecedent. In ergative case marker there are few bindings and exceptions, e.g. the postposition "ne" (ने) must come right after the subject; the subject changes in oblique the perfect form of the verb now agrees with the direct object in number and gender. In case of above condition, number and gender agreement puts the bindings between verb and direct object. Exceptions are:

- If the object is not stated, or if the object is followed by को (ko) then the perfect form of the verb should be in masculine singular form.
- The auxiliary verb (if any) also agrees with the object, not the subject.

The eight case markers are ergative, nominative, ablative, accusative, instrumental, genitive, dative and locative. We have considered these eight cases of Hindi for the study. Few cases are given below, though in linguistic perspective, these cases come as a suffix.

- ne – it marked as ergative marks the subject or topic (but only in the past perfective tense for transitive verbs)
- ka/ke/ki - marks the genitive
- ko - marks the accusative or dative, - typically means "from" or "by", also marks the passive agent
- mein, par - locative "in", "at"

E. *Accuracy measurement of automated tags*

In natural language processing mainly two metrics are used. The First precision and the second recall. Another metrics is derived from precision and recall, which is called F measure

- Precision (P) is the fraction of retrieved documents that are relevant

$$p = \frac{\#(relevant\ item\ retrieved)}{\#(retrieved\ items)} \quad (1)$$

- Recall (R) is the fraction of relevant documents that are retrieved

$$R = \frac{\#(relevant\ item\ retrieved)}{\#(relevant\ item\ )} \quad (2)$$

These notions can be made clearer by examining the following contingency table:

TABLE II. CONTIGENCY TABLE

	Relevant	**Non relevant**
Retrieved	True positive(tp)	False positive(fp)
Not retrieved	False negative(fn)	True negative(tn)

$$P = \frac{tp}{(tp+tf)} \quad (3)$$

$$R = \frac{tp}{(tp+fn)} \quad (4)$$

There is another alternative to calculate accuracy of data set. It is the ration of true selection of text and the sum of all selections (all true + all false).

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn} \quad (5)$$

This seems plausible, since there are two actual classes, true and false, and an information retrieval system can be considered as a two-class classifier which attempts to label them as such. We are using only equation first, second and seventh.

- F measure : It is the harmonic mean of Precision and Recall

Given n points, x1, x2,…………x_n, the harmonic mean is:

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n} 1/x_i \quad (6)$$

So, the harmonic means of precision and recall is:

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right) = \frac{P+R}{2PR} \quad (7)$$

With the help of above equation (7), we will calculate the F measure. The authenticity of results checked against only three metrics, i.e. precision, recall and F measure.

VI. Result and Discussion

The study was conducted on twelve files of Ranchi Express from EMILEE corpus. It consists 206 news items, and 1485 demonstrative pronouns. An effort has been made to automate tagging of last three tags given in table 1. A set of rules are applied to accomplish the task. Three accuracy metrics have been considered; precision, recall and F measure. The Table 3 shows all the twelve files and number of pronouns found in the respective file. Number of news and length of news per file is directly proportional to the number of pronouns.

TABLE III. Pronoun Count Per File

File No.	Number of Pronoun / file
1.	61
2.	148
3.	101
4.	108
5.	104
6.	155
7.	138
8.	128
9.	102
10.	111
11.	151
12.	178
Total	1485

In each file for each pronoun performance metrics are calculated. In the first file, first pronoun was “iss” (इस) and the value of precision, recall & F measure is 66.66, 66.66 and 66.66 respectively. These values for second pronoun are 100, 100 and 100. The average precision, recall and F measure of file 1 at serial number one in the table. Hence, the average of each metric is calculated and given in the table 4 below along with its respective file. Then again average of all files is calculated. Precision varies from 65.52 to 79.08 and recall 65 to 75. This depicts there is no major variation in all metrics.

In figure 1, the annotated genre; on the axis X number of files are given. On axis Y, percentage of accuracy of three tags in the terms of precision, recall and F measure is given. Though the pattern of result of all these three parameters for all the twelve files is almost the same. The average accuracy of precision, recall and F measure is 74, 71 and 72 (approximately) respectively.

A. Error Analysis

We classify the entire results into three types; worst, average and best results across the data set (all files) and their pronoun.

TABLE IV. Precesion, Recall and F Measure Per File

S. No.	File Name	Precision	Recall	F measure
1.	File1	77.22	75.66	76.3
2.	File2	78.83	75.30	76.79
3.	File3	66.81	71.61	68.82
4.	File4	76.94	73.44	74.84
5.	File5	78.96	74.98	76.67
6.	File6	74.24	70.22	71.93
7.	File8	68.51	65.02	66.44
8.	File9	72.70	68.34	70.11
9.	File10	76.89	72.24	74.17
10.	File11	79.08	75.53	76.97
11.	File12	73.70	69.20	70.96
12.	File13	65.52	61.94	63.30
Average of each metrics		**74.12**	**71.12**	**72.28**

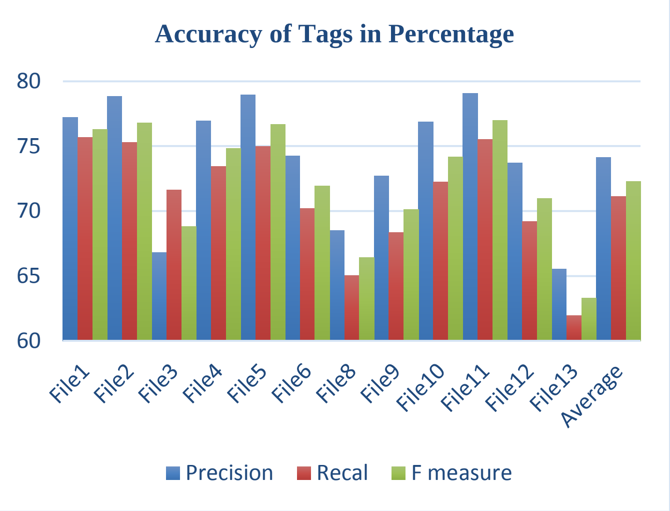


Fig. 1. Accuracy of tags

*1) **Worst case (file 8, pronoun 81st):*** It is observed that in the 7th file (file name is file 8) having 81st to 83rd pronoun produced the result zero for precision, recall and F measure**.** Automated file has been given below. Special cases, which annotate genre with zero accuracy compared with manual tagged file for the same text with same pronoun. In the proposed algorithm, first, we stored all case marker/ connector and pronoun patterns. The algorithm search for the required case marker and pronoun pattern. Wherein window size is more than one in which we are seeking the case marker and the pronoun patterns.The precision and recall was calculated of the entire string after the 7th tag. In manual tagging there are four entries, including pronoun after the 7th tag. On the other hand automatically tagged 81st pronoun has ten entries after the 7th tag. Last three tags are same in automatic annotation and manually tagged annotation for this example.

Pronoun 81st : अगर वास्तव में <w tag="D,D,D,A,R,M,N,उन,इन,उन्होंने,उनका,वे,उसी,उन,_,_,_">उन</w> लोगों ने छेड़छाड़ की घटना के मुद्दे पर ही

Pronoun 82nd : <w tag="P,D,D,A,R,M,N,इन,उन्होंने,उनका,वे,उसी,उन,_,_,_">इन</w> सिख युवकों की हत्या की थी तो निश्चित रूप से

Pronoun 83rd : <w tag="D,P,D,A,R,H,N,उन्होंने,उनका,वे,उसी,उन,_,_,_">उन्होंने</w> योजना बनाकर <w

The pronoun 82nd and 83rd of the same file depict the same result due to over length entries in string of case marker and pronoun pattern. It can be fixed with the help of window size. File 3, file 8, file 12 falls in the category of the worst case. Though the average of these files is 61 to 75 percentage.

Manually tagged pronoun for comparison of file 8.

Pronoun 81st : अगर वास्तव में <w tag="D,D,D,A,R,M,N,un,null,null,null">उन</w> लोगों ने छेड़छाड़ की घटना के मुद्दे पर ही

Pronoun 82nd <w tag="P,D,D,A,R,M,N,inn,null,null,null">इन</w> सिख युवकों की हत्या की थी तो निश्चित रूप से

Pronoun 83rd : <w tag="D,P,D,A,R,H,N,unhon-ne,null,null,null">उन्होंने</w> योजना बनाकर

2) **Average case (file 1, pronoun 1st):** It was considered as average case if the accuracy of precision, recall and F measure is fifty percent or more. File 1, file 4, file 6, file 8, file 9 and file 11 lies in the average case.

Manually annotated file1

<body>

<p>बालूमाथ, १६ मार्च: आज शाम बालूमाथ थाना अंतर्गत लेबडाही जंगल में प्रतिबंधित एम.सी.सी. संगठन के उग्रवादियों और पुलिस के बीच मुठभेड़ हुई।

Pronoun 1st : <w tag="P,D,D,A,R,M,N,iss,null,null,act">इस</w> मुठभेड़

Automatic tagged file1

<body>

<p>बालूमाथ, १६ मार्च: आज शाम बालूमाथ थाना अंतर्गत लेबडाही जंगल में प्रतिबंधित एम.सी.सी. संगठन के उग्रवादियों और पुलिस के बीच मुठभेड़ हुई।

Pronoun 1st : <w tag="P,D,D,A,R,M,N,इस,_,_,_">इस</w> मुठभेड़

In the above example of file 1 and pronoun 1st, only one is/has mismatched tag. Tag 10 differs in manual and automatic tagging.

3) **Best Case (file 1, pronoun 2nd):** It was considered as the best case if all three metrics have 100 percent accuracy. In this example null is equal to _ (underscore) sign. Each field matched and in both files.

Manually annotated file1

Pronoun 2nd : <w tag="P,D,D,A,R,M,C,iss,null,null,null">इस</w> गोलीबारी में पुलिस दस उग्रवादियों को मार गिराने का दावा कर रही है। उग्रवादियों की गोली से बालूमाथ थाना पुलिस वाहन के चालक अशोक कुमार (३५) गंभीर रूप से घायल हो गया है। पुलिस और एम.सी.सी. के बीच ढाई घंटे तक मुठभेड़ हुई है।

Automatic tagged file1

Pronoun 2nd : <w tag= "P,D,D,A,R, M, C, इस, _,_,_">इस</w> गोलीबारी में पुलिस दस उग्रवादियों को मार गिराने का दावा कर रही है। उग्रवादियों की गोली से बालूमाथ थाना पुलिस वाहन के चालक अशोक कुमार (३५) गंभीर रूप से घायल हो गया है। पुलिस और एम.सी.सी. के बीच ढाई घंटे तक मुठभेड़ हुई है।

B. *Inference:*

- It reflects that the seeking window should be decreased to the optimal size in order to avoid additional case marker and pronoun pattern. This particular example (**Worst case, file 8, pronoun 81st**) depicts that the scenario of case marker string and pronoun pattern string may have multiple entries due to the number of pronoun and the number of case marker come consecutively in discourse. e.g. उन (un), इन(in), उन्होंने(unhone).

- The last three features was automated. There are two methods to match ten features. First, start the matching of features from 7th tag to 10th tag. And the second is to match the last three features of automatic and manual files. Then discard the additional entries between 7th and last but 3rd. It will solve the problem of additional entries in case marker and pronoun string. It also will improve the accuracy in the terms of precision, recall and F measure.

- Refining the number of rules will increase the accuracy of automatic annotation. These rules define the pattern of case marker/ connectives and pronoun. These two features have more contribution for error.

VII. Conclusion and Future Work

A. *Depiction of results*

Few results were concluded from ongoing work. Before arriving at the conclusion, twelve files are studied of monologue, and 206 news items which consist 1485 pronouns. Accuracy varies from 65 to 79 for precision. Recall varies 62 to 75 and F measure has been maximum and the minimum values are 77 and 63 percent respectively. Average of precision, recall and F measure remained 74, 71 and 72 percent. File number twelve (12) has the lowest accuracy for all the three metrics and file ten (10) has the highest.

B. *Conclusion*

This dataset depicts the generalized picture of the genre. Fine tuning of rules besides considering few other semantic categories increase the accuracy of results. In the proposed algorithm the window size is five or till the end of sentence. It will count all the pronoun patterns and add to the tag set. It will make additional entries in values of feature set string. The remedial action for window size is to decrease it to one. Now, there are few advantages and few disadvantages of window size one.

Decreasing window size may increase the accuracy in terms of precision, recall and F measure. While the disadvantage is that it will skip further potential pronoun pattern which may yield higher accuracy. That means we have to go for the best hit. To find the best hit one has to develop new logic. Potential pronoun pattern and case marker may be tested for other genres. It also optimizes the results. In this work we are replacing the potential pronoun pattern with the first hit. In future it may be replaced with the best hit by other logic. Fine tuning the rules and taking gold standard dataset increase the percentage of accuracy.

C. Future work

Though the corpus is small. A larger and gold standard dataset may produce more authentic results. Selection and fine tuning of rules increase the accuracy of results. Developing the logic for the best hit in seeking window for potential pronoun pattern and case markers also helps to improve the results. These few issues can be addressed in future work which optimize the results.

ACKNOWLEDGMENT

The author would like to acknowledge Pratheek Adidela for discussion on python libraries, their use and detailed discussion on the logic of indirect anaphora and the sequence of rules to be applied in the given algorithm.

REFERENCES

[1] M. Marcus, B. Santorini and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," Association for Computational Linguistics, pp. 313-330, 1993.

[2] G. Sampson, English for the computer, The SUSANNE corpus and analytic scheme, Oxford, UK: Clarendon Press, 1995.

[3] G. Leech, "The Lancaster Parsed Corpus," ICAME Journal, vol. 16, no. 124, 1992.

[4] S. Greenbaum, Comparing English worldwide: The International Corpus of English, Oxford, UK: Clarendon Press, 1996.

[5] J. Hajic, "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank," in Issues of valency and meaning, Karolinum, Praha, 1998, pp. 106-132.

[6] A. Abeill´e, L. Cl´ement and A. Kinyon, "Building a treebank for french," in Proceedings of the Second International Conference on Language , Athens, Greece., 2000.

[7] C. Bosco, V. Lombardo, D. Vassallo and L. Lesmo, "Building a treebank for italian: A data-driven annotation schema," in Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Greece, 2000.

[8] A. Moreno, R. Grishman, S. L´opez, F. S´anchez and S. Sekine, "A treebank of spanish and its application to parsing," in Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Greece, 2000.

[9] K. Oflazer, D. Hakkani-T¨ur and G. T¨ur, "Design for a turkish treebank," in proceedings of the Workshop on Linguistically Interpreted Corpora LINC-99, Bergen, Norway, 1999.

[10] I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin and N. Frid, "Dependency treebank for russian: Concept, tools, types of information," in 18th International Conference on Computational Linguistics COLING-2000, Saarbr¨ucken, Germany, 2000.

[11] S. Brants and S. Hansen, "Developments in the TIGER Annotation Scheme and their Realization in the Corpus," in proceedings of the Third Conference on Language Resources and Evaluation LREC-02, 2002.

[12] P. James, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer and D. Radev, "The timebank corpus," Corpus linguistics, pp. 40-48, 2003.

[13] G. Leech and R. Garside, "Running a Grammar Factory: The Production of Syntactically Analysed Corpora or Treebanks," English Computer Corpora: Selected Papers and Research Guide, pp. 15-32, 1991.

[14] R. Mitkov, R. Evans, C. Orasan, C. Barbu, L. Jones and V. Sotirova, "Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies," in Proceedings of DAARC 2000, UK, 2000.

[15] A. Tutin, F. Trouilleux, C. Clouzot, E. Gaussier, A. Zaenen, S. Rayot and G. Antoniadis, "Annotating a large corpus with anaphoric links," in Third International Conference on Discourse Anaphora and Anaphor Resolution DAARC2000, United Kingdom, 2000.

[16] A. P. Belis, L. Rigoust, S. Salmon-Alt and L. R, "Online Evaluation of Coreference Resolution," in LREC 2004 Fourth International Conference on Language Resources and Evaluation, 2004.

[17] S. M. Harabagiu and S. J. Maiorano, "Multilingual coreference resolution," in Proceedings of the sixth conference on Applied natural language processing, Morristown, NJ, USA, 2000.

[18] M. D. Swift, M. O. Dzikovska, J. R. Tetreault and J. F. Allen, "Semi-automatic syntactic and semanticcorpus annotation with a deep parser," in Fourth International Conference on Language Resources and Evaluation LREC-2004, 2004.

[19] Y. Esteve, T. Bazillon, J. Y. Antoine, F. Bechet and J. Farinas, "The EPAC corpus: Manual and automatic annotations of conversational speech in french broadcast news," in proceedings of the seventh conference on International Language Resources and Evaluation(ELRA), Valletta, Malta, 1686-1689.

[20] E. Hinrichs and T. Zastrow, "Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012.

[21] M. Palmer, D. Gildea and P. Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Role," Computational Linguistics archive, vol. 31, no. 1, pp. 71-106, 2005.

[22] S. P. Botley and A. McEnery, "Demonstratives in English: a corpus-based study," Journal of English Linguistics, vol. 29, pp. 7-33, 2001.

[23] S. P. Botley, "Indirect anaphora: Testing the limits of corpus-based linguistics," International Journal of Corpus Linguistics, vol. 11, pp. 73-112, 2006.

[24] M. Halliday and R. Hasan, Cohesion in English, London: Longman, 1976.

[25] S. Sinha, "A Corpus-based Account of Anaphor Resolution in Hindi," UK, 2002.

[26] R. Prasaad, E. Miltaski, A. Joshi and B. Webber, "Annotation and Data Mining of the Penn Discourse TreeBank," in ACL Workshop on Discourse Annotation, 2004.

[27] K. Dutta, S. Kaushik and N. Prakash, "Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items," The Prague Bulletin of Mathematical Linguistics, pp. 33-50, 2011.

An Effective Storage Mechanism for High Performance Computing (HPC)

Fatima El Jamiy, Abderrahmane Daif, Mohamed Azouazi and Abdelaziz Marzak
University Hassan II Mohammedia, Faculty of Sciences Ben M'sik,
Laboratoire Mathématiques Informatique et Traitement de l'Information MITI, Casablanca, Morocco

***Abstract*—All over the process of treating data on HPC Systems, parallel file systems play a significant role. With more and more applications, the need for high performance Input-Output is rising. Different possibilities exist: General Parallel File System, cluster file systems and virtual parallel file system (PVFS) are the most important ones. However, these parallel file systems use pattern and model access less effective such as POSIX semantics (A family of technical standards emerged from a project to standardize programming interfaces software designed to operate on variant UNIX operating system.), which forces the MPI-IO implementations to use inefficient techniques based on locks. To avoid this synchronization in these techniques, we ensure that the use of a versioning-based file system is much more effective.**

Keywords—Big data; High Performance Computing; Storage; Distributed File System; BlobSeer

I. Introduction

Industrial research and development on parallel file systems that can provide outstanding performance is prompted by the need to treat raising volume of data in technological and business applications that usually require high Input/output throughput [1]. Among these, we can cite Physical Simulation, processing a big volume of data sets to extract knowledge and business email services. In this article, we will present two important parallel file systems while addressing their major limitations and propose a new File System based on versioning and inspired from PVFS and Lustre File systems. Our choice of these parallel file systems is mainly due to their extensive use. Although both systems have many differences in their design such as Locking, Semantics, Caching and Striping Pattern, they have the same fundamentals of Striping Width and metadata management. The main purpose of this document is to emphasize the strengths of the two systems and present a prototype of a new file system based on the BlobSeer storage system that provides high Input/output throughput while ensuring simultaneous access data for distributed file systems.

II. Main Problem and Approach

HPC is traditionally defined by parallel scientific applications that are becoming more and more intensive on data and whose I/O (input-output) performances become quickly a problem, causing a bottleneck that has a negative impact on the overall application performance [2].

It has been found that most software components of parallel computing systems are in place such as operating systems, local storage systems, and message passing systems. However, one area is devoid of components to the production level for clusters, it is the one of systems parallel I/O [3].

The HPC system architecture is divided into several support layers (Figure 1) that provide much functionality:

- Abstractions to data structure, cell (eg netCDF parallel, Adios)

- Manage the organization of the data access

- Sustain a logical space (eg Lustre and PVFS); it manages the storage hardware and provides a single view, while focusing on simultaneous and independent access.

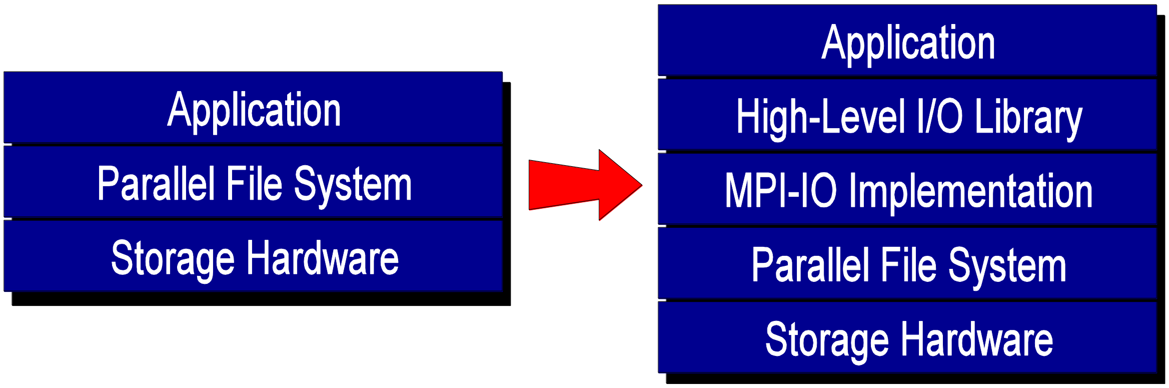


Fig. 1. High Performance Computing Architecture

A. PVFS and LUSTRE

PVFS is a scalable high performance parallel file system for HPC systems. Distributing data through several nodes named Input/Output nodes is the main method used to offer a high access to data kept in the file system by numerous users. The expansion of data that way enable applications to use many directions to access data and thus eliminate bottlenecks while the total throughput is also improved [4]. Although the traditional mechanism of access to the file is convenient and allows all applications to access stored files on many different types of file systems, there is an overload in access through the core [5].

B. LUSTRE

Lustre is an open source project (GPL) distributed file system based on objects. Its name is the mixture of the words "Linux" and "cluster". The system architecture consists of metadata server (MDS), the target object storage servers (OST) and customers. The metadata servers maintain data about all files in Lustre [6].

In fact, to guarantee the atomicity of access to non-contiguous and overlapping areas, parallel file systems use access patterns less powerful and less effective such as POSIX semantics that forces MPI-IO Implementations (Message Passing Interface) to use inefficient techniques based on locks [7]. To avoid this synchronization in these techniques, we ensure that the use of a versioning-based file system is much more effective. Our prototype will be greatly inspired from PVFS and Lustre and It will be mainly an evolution of PVFS, it will re-take the same principles of operation (cell, volume, persistent caching, replication,). Its first goal is to provide high Input/output throughput while providing simultaneous data access.

C. Blobseer Architecture

BlobSeer is a large-range distributed storage service that meets the advanced management data from a large mass of data requirements [8]. It is based on the use of versions to manipulate simultaneously a large binary objects (BLOBs) in order to effectively exploit parallelism in data and sustain a high throughput despite the massively data access parallel [9] .

The client controlled and handled the Versioning in the system where Each BLOB (Binary large object) has its own unique key. When writing or adding, the data to be written is divided into a number of small pieces and are written in the data providers listed by the provider vendor manager [10]. New versions are then produced in each write or append, however storage space is conserved because what is kept is just the dissimilar patch. During a read operation, the latest version number is first obtained from the manager version. All pieces that match this version are identified by the client and then perform parallel read operation [11]. BSFS does not have master-slave architecture and thus released from the single point of failure. The biggest benefit is that metadata are highly distributed between metadata providers. So there is no fear of point failures that will stop the availability of metadata and slow down all operations that depend on it. Another feature that adds to the attraction of this architecture is the versioning technique. Indeed, the amount of treatments accomplished in the parallel file system is optimized because the concurrency control algorithm [12].

D. Approach:

The principal point where Lustre and PVFS have differences is in the method they use to split the metadata. In fact the metadata management is an important element in offering scalability and performance, and in this context PVFS does not give any assurance that quality will be fulfilled since all the tasks related to that are distributed across the servers' without taking in consideration that factor. Whereas on the other side, Lustre reach an elevated availability but it does not enhance performance. To attain this, two servers are used in a consolidation plan.

When concurrent requests occurred, constancy and stability are provided by Lustre, whereas PVFS does not bear that implementation plan, this possible only if we have non overlapping areas to access. PVFS does not provide POSIX semantics [14]. The atomicity of writes is guaranteed in non-overlapping areas and even in non-overlapping, non-contiguous regions. It does not implement a lock infrastructure [15].

The atomic mode ensures that data written to a process is immediately visible to another process (like POSIX semantics by default). ROMIO currently uses file locking to implement the functionality of the atomic mode MPI-IO [16]. Locks in PVFS are not supported and therefore the atomic mode is not supported too [17, 18, 19].

The implementation of PVFS does not include all the features of MPI-IO specification. Eventually we reached a point where it was obvious to us that a new design is required. Our conception embodies the principles of MPI-IO components missing for PVFS that we consider key to an efficient, robust and high performance parallel file system.

The efficient design oriented version of BlobSeer enables a lock-free data access, and thus promotes scalability under high concurrency. A high I/O debit data is offered by Blobseer because of his particular characteristics and decentralized data and metadata management. This realization aim to come up with a new vision that demonstrate the way Blobseer can be employed as effective backend storage by expanding it to a distributed file system for HPC systems.

We will configure our new file system and evaluate its performance against the PVFS performance and Lustre on a set of Data-intensive computing benchmarks and real systems.

III. Comparison with Related Work

There have been many works aiming to develop and enhance the performance of parallel systems but until now they cannot bring the results sought by this kind of systems. Hadoop, an open source framework designed to carry out processing on massive data volumes, on the order of several petabytes (or several thousand TB) and written in Java has been improved with PVFS. PVFS is a widely used parallel file system that allows a high performance data access for the operations I / O that are adjacent and non-contiguous without guaranteeing atomicity.

Continuing with Hadoop, another experiment has been done to enhance it. It was integrated with Blobseer to allow a high access data and avoid synchronization but it is still not enough to take over all the requirements.

In another work, the authors propose a lock pattern for a non-contiguous access strictly aimed at reducing the scope of the locked region to areas that are really accessed. However, this approach does not prevent the serialization for the overlapping and simultaneous Input/Output.

IV. Conclusion

Our work confirm that using a new layer created with the different principles of conception cited above is apt to improve the efficiency of data storage layer and thus that of the whole HPC applications. With the new layer of the file system BlobSeer (BSFS), we will propose a new file system inspired from the two principal distributed file systems Lustre and PVFS. The next step is the implementation of this new file system on the Grid'5000 infrastructure and evaluates its performance.

References

[1] J. Dean and S. Ghemawat, “MapReduce: Simpli_ed Data Processing On Large Clusters, Commun. ACM, vol. 51, 2008.

[2] Mahadev Satyanarayanan, James J. Kistler, Puneet Kumar, Maria E. Okasaki, Ellen H. Siegel, and David C.Steere, “Coda: A highly available file system for a distributed workstation environment,” IEEE Trans. Comput., 1990.

[3] A. Ching, K. Coloma, and A. Coudhary, Challenges for Parallel I/O in GRID Computing. Publisher’s address: American Scientific Publisher, 2006, ch. 6, Grid I/O.

[4] K. Shvachko, “Hdfs scalability: The limits to growth”, The USENIX Magazine , 2010.

[5] J. Dean and S. Ghemawat. MapReduce: A Flexible Data ProcessingTool. CACM, 53(1):72–77, 2010.

[6] S. Ghemawat, H. Gobioff, and S.-T. Leung. The Google file system.In SOSP, 2003.

[7] C. Yan, X. Yang, Z. Yu, M. Li, X. Li, IncMR: incremental data processing based on MapReduce, in: Proc. Int’l Conf. Cloud Computing, CLOUD, 2012.

[8] S. Wu, F. Li, S. Mehrotra, and B. C. Ooi. Query Optimization for Massively Parallel Data Processing. In SOCC, 2011.

[9] S. Babu. Towards automatic optimization of MapReduce programs. In SOCC, 2010.

[10] J. Shafer, S. Rixner, and A.L. Cox, “The hadoop distributed filesystem: Balancing portability and performance”,IEEE international symposium, 2010.

[11] J. Dittrich, J.-A. Quian´e-Ruiz, S. Richter, S. Schuh, A. Jindal, and J. Schad. Only Aggressive Elephants are Fast Elephants. PVLDB, 2012.

[12] William Gropp, Ewing Lusk, and Anthony Skjellum. Using MPI : Portable Parallel Programming with the Message Passing Interface. MIT Press, 1994.

[13] Peter Aarestad, Avery Ching, George Thiruvathukal, and Alok Choudhary. Scalable approaches for supporting MPI-IO atomicity. In Proceedings of the IEEE/ACM International Symposium on Cluster Computing and the Grid,May 2006.

[14] Randal E. Bryant, Data-intensive scalable computing for scientific applications, Comput. Sci., 2011.

[15] S. Weil, S. Brandt, E. Miller, D. Long, C. Maltzahn, “Ceph: A Scalable, High Performance Distributed FileSystem,” In Proc. of the 7th Symposium on Operating Systems Design and Implementation, Seattle, WA, 2006.

[16] Avery Ching, Alok Choudhary, Kenin Coloma, Wei Keng Liao, Robert Ross,and William Gropp. Noncontiguous access through MPI-IO. In Proceedings of the IEEE/ACM International Symposium on Cluster Computing and the Grid, 2003.

[17] Philip H. Carns, Walter B. Ligon III, Robert B. Ross, and Rajeev Thakur. PVFS: A parallel file system for Linux clusters. In Proceedings of the 4th Annual Linux Showcase and Conference, pages 317–327, Atlanta, GA, 2000. USENIX Association.

[18] Avery Ching, Alok Choudhary, Wei-keng Liao, Rob Ross, and William Gropp. Noncontiguous I/O through PVFS. In CLUSTER ’02 : Proceedings of the IEEE International Conference on Cluster Computing,CLUSTER ’02, pages 405–,Washington, DC, USA, 2002. IEEE Computer Society.

[19] K. Shvachko, H. Huang, S. Radia, and R. Chansler, “ The hadoop distributed file system,” In 26th IEEE (MSST2010) Symposium on Massive Storage Systems and Technologies, 2010.

SLA for E-Learning System Based on Cloud Computing

Doaa Elmatary
Industrial Technical Institute
Technical college, Port Said
Port said, Egypt

Wael Awad
Department of Mathematics and Computer Science
Faculty of Science, Port Said University
Port Said, Egypt

Samy Abd El Hafeez
Department of Mathematics and Computer Science
Faculty of Science, Port Said University
Port Said, Egypt

Fatma Omara
Department of Computer Science,
Faculty of Computers and Information, Cairo University,
Cairo, Egypt

Abstract—**The Service Level Agreement (SLA) becomes an important issue especially over the Cloud Computing and online services that based on the 'pay-as-you-use' fashion. Establishing the Service level agreements (SLAs), which can be defined as a negotiation between the service provider and the user, is needed for many types of current applications as the E-Learning systems. The work in this paper presents an idea of optimizing the SLA parameters to serve any E-Learning system over the Cloud Computing platform, with defining the negotiation process, the suitable frame work, and the sequence diagram to accommodate the E-Learning systems.**

Keywords—Cloud Computing; Service Level Agreement (SLA); E-learning System

I. Introduction

The Cloud Computing defined as delivering the hosted services over the Internet in an on-demand model. A cloud provider offers resources (e.g., hardware, software or development stacks) as services over the internet depending on a pay-as-you-use basis [1]. In the Cloud Computing, the users could access applications and services by using his browser no matter where these applications and services are hosted [1]. Using the applications of the Cloud Computing provides the user an effective way to use the infrastructure, and to maintain his services where he does not need to waste his time in installing and maintaining any components on his device [2].

The Cloud Computing Models can be divided into three layers, every layer handle a certain service; Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) (see Figure [1]) [3]. Software as a service (SaaS) eliminates the need for installing and running the application on the local devices of users, so that the users can access the network and the available programs remotely via internet [2]. Platform as a service (PaaS) provides the necessary framework to develop the applications without the cost and complexity of buying and managing the underlying hardware [3]. Infrastructure as a service (IaaS) offers the required physical infrastructure (e.g., servers, storage, processors) [2].

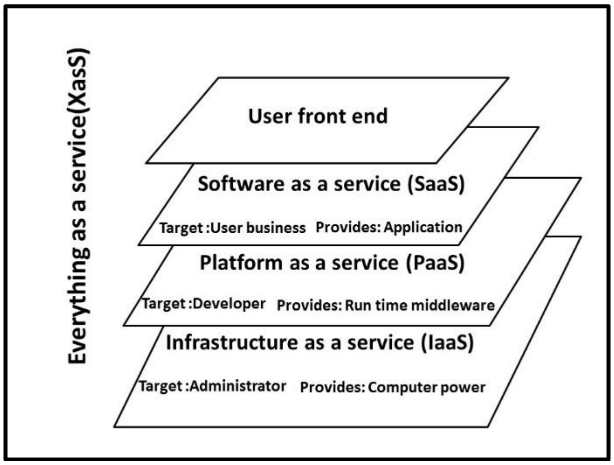

Everything as a service(XasS)
User front end
Software as a service (SaaS)
Target :User business
Provides: Application
Platform as a service (PaaS)
Target :Developer
Provides: Run time middleware
Infrastructure as a service (IaaS)
Target :Administrator
Provides: Computer power

Fig. 1. Architecture of Cloud Computing Layers [4]

One of the main aspects in the Cloud Computing is isolating the user from the internal details of the system, which means that there is no control from the user on the computing resources. Thus, the user needs to guarantee the resources availability and reliability, this guarantee is provided through a negotiation between the provider and consumer to create a contract with some criteria called Service Level Agreements (SLAs) [2].

On the other hand, the SLA is a contract between the consumer and the provider. The main issue of SLAs is to provide a clear definition of the formal agreements about the service terms like performance, availability and billing. It is important that the SLA includes the obligations and actions that can be taken in the case of any violation, with clearly defined semantics between each party involved in the online contract [6].

As the SLA becomes an important issue for any user over the cloud, it can be used for different kinds of applications (e.g., medical, commercial and learning systems). According to the work in this paper, the SLA is applied on the E-Learning system that serves a different kind of users.

II. Related Works

Selecting a Template (Heading 2) The SLA has been used since the late 1980s by telecom operators as a part of the contract with their customers. After that, the SLA becomes a standard protocol of business applications and Web Services [9]. Generally, There are two main specifications are designed to describe the SLA; 1) The Service Negotiation and Acquisition Protocol (SNAP) which support reliable management of remote SLAs and describe the negotiating process in the system [10]. 2) The conceptual SLA frame work for Cloud Computing that describes the main characteristics of SLAs in Cloud Computing and explains the SLA parameters specified by metrics for the four types of cloud services (i.e., IaaS, SaaS, PaaS, Storage as a Service) [6]. Chenkang,W.,et. al [11] have suggested different implementation structure of SLAs parameters by adding another dimension related to both of the providers and the users. Concerning the IaaS, the SLAs represent the metrics of the performance for the provider, and the experience for the users. While in the PaaS, both providers and users need more guarantees for the integration's ability and scalability. Finally, for SaaS, the providers need to offer some specialized functions like multi-terminal supporting and customization, and the users should care about the stable usability.

Patel, P., et. al. [12] has proposed a framework with a main concept and architecture for the Web Service Level Agree (WSLA). It is a contract defines the web services where the cloud provider is out of their scope.

All of the previous researches handled the SLA in general without applying its principles on a specific application by describing how the cloud services are offered to the customers using SLA (Service Level Agreement).

On the other hand, Shrivastava, G. [13] has applied the Cloud Computing to serve the E- Learning system. The availability of online resources in the Cloud Computing gives the student the opportunity to share the E-Learning resources and access it online at anytime and anywhere.

Finally, this research is based on the idea of optimizing the SLA parameters, negotiation process, and the frame work to accommodate the E-Learning systems. Building the E-Learning server on Cloud Computing gives the ability to adjust the required resources as needed, and allows the users to get the benefits of Cloud Computing services.

III. The Service Level Agreements (SLAs)

The SLA is a legal format to document the way of delivering the service, as well as, providing a framework for paying service charges. Both the providers and users dealwith SLAs from their points of view. From the Service provider's side, he uses the contract to optimize his use of infrastructure to meet the signed terms of the services. While, the Service user deals with the SLA to ensure the required level of service quality and to maintain acceptable business models for long-term provisioning of services [6]. The main requirements of the SLAs can be concluded as follows [6]:

- SLA's format should clearly describe a service so that the service's user can easily understand the operations that can be done on the services.

- SLA has to present the level of performance of service.

- SLA has to define the ways of monitoring the service parameters and the format of reports monitoring.

- SLA has to clarify the penalties when the service requirements are not met.

- SLA presents the business metrics such as billing values when this service is done without any penalties.

The SLA parameters are specified by metrics; these metrics define how the parameters of the cloud service can be measured. Usually, these metrics are varied from an application to another. So in this paper, the SLA parameters are specified only for the E-Learning applications [6]. Most users are confused in defining the important parameters. For the E-Learning applications, four types of services which providers can provide to the users. These services are IaaS, SaaS, PaaS, and Storage as a Service [6]. For each part of the SLA, the most important parameters that the users can use to create a reliable model of negotiation with the service provider are defined [6].

This paper focuses on the most influence parameters that will serve any E-Learning system. Where the E-Learning system will be created based on the Cloud Computing, thus all the required resources can be adjusted as needed.

Table (1) represents the SLA metrics of SaaS for the E-Learning system

TABLE I. SLA Metrics For SaaS

Parameter	Description for the Educational Service
Reliability	The system can be still working in most cases.
Usability	Easy and Clear Interface
Scalability	Using with one user only or with all users.
Customizability	Flexible to use with different type of users.

Table (2) represents one of the most important parameters in the SLA metrics of IaaS for the E-Learning system, where a sufficient area (storage) has to be created to keep the required data of the system whether for short or long contract agreement. This can be done by reserving a storage unit from the Cloud provider.

TABLE II. SLA Metrics for IaaS

Parameter	Description for the Educational Service
Storage	Storage "Z" (big) size of data for the organization. Storage "Z*" (small) size of data for the student.

Table (3) represents the SLA metrics of Storage as a Service that should be a guarantee for the E-Learning system.

PaaS can be provided to extend the E-Learning system by offering extra services to all users. So, the software applications have to access the platform by using the user defined browsers [8]. The SLA metrics of PaaS of E-Learning system are displayed in Table (4).

TABLE III. SLA METRICS FOR STORAGE AS A SERVICE

Parameter	Description for the Educational Service
Storage billing	For the price billing, the most influence units are the processing and storage.
Security	No one can access the material without registration and authentication. That is to prevent the system from the illegal access.
Privacy	(validation ,authentication ,verification)
Backup	Images of data (database of the users' profiles) are stored according to the regulation of the E-learning system, and users have the ability of recovery the profile data in case of disaster occurrence.

TABLE IV. SLA METRICS OF PAAS FOR THE E-LEARNING SYSTEM

Parameter	Description for the Educational Service
Integration	Coordination with other institutions to extend the platform.
Scalability	Extends itself so estimated offering services to all users.
Pay as you go billing	Pay only for the service that you use.
Browsers	User defined

IV. THE PROPOSED SLA FOR E-LEARNING SYSTEM

The proposed SLA for E-Learning system represents the metrics should be defined for the "E-Learning system" over the Cloud Computing. At this moment, the SLA negotiation process between the students and the instructors is needed to be explained. The frame work of the E-Learning's SLA is shown in Figure (2). It consists of three entities; **Parties, Service definition, and Obligation Services.**

A. Parties

It consists of the parties of the agreement. The service's provider deals with the coordinator of the E-Learning system who is considered here the service user. Then, both the student and the instructor deal with the coordinator, where he is considered the service provider. So, the coordinator could be considered as the service broker.

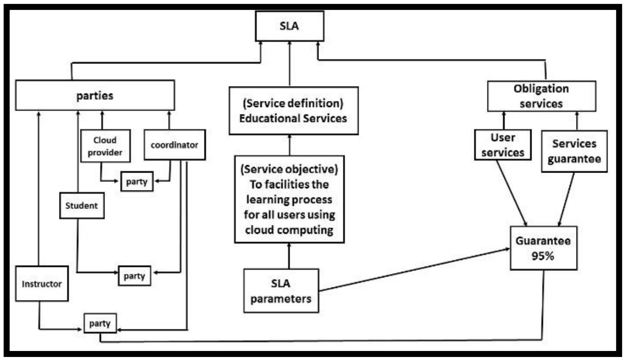


Fig. 2. Frame Work of SLA for E-Learning System

B. Service Definition

It defines the educational service and its objective to improve the facilities of the learning process for the users (i.e. students or instructors) by using the Cloud Computing environment. This component contains SLA parameters for the service. As shown in Figure (3), it has two main cloud metrics (i.e. infrastructure as a service and Storage as a service).

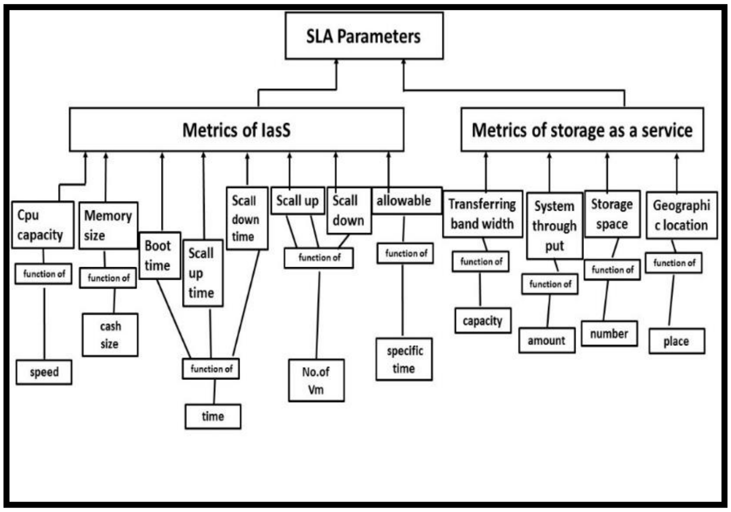


Fig. 3. SLA of Service Definition

C. Obligation Services

It contains the service guarantee represented in common parameters for all of the users (students and instructors) to ensure the services that are delivered (see Figure (4)).

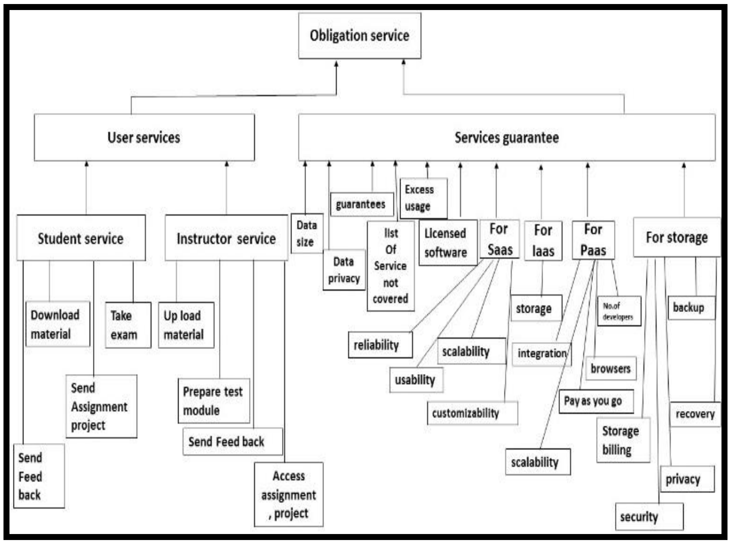


Fig. 4. Obligation Services of SLA for E-Learning System

V. SLA RELATIONAL DIAGRAM FOR THE E-LEARNING SYSTEM

Relational diagram (RD) is a graphical representation of an information system that shows the relationship between people, objects, places, concepts or events within that system [14].

The three main components of the RD are; 1) the entities, which are objects or concepts that store the information, 2) the relationship between those entities, and 3) the cardinality, which defines that relationships in terms of numbers [14]. The proposed Relational Diagram consists of four components; cloud provider, negotiation process, system implementation, and SLA document (see Figure (5)).

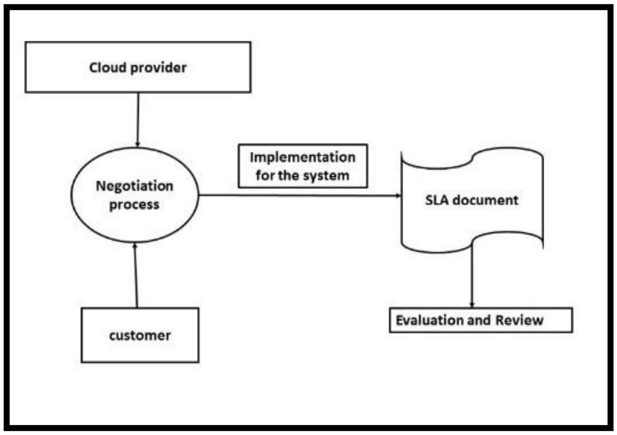


Fig. 5. SLA Relational Diagram for E-Learning System

A. Negotiation Process

Between the service's provider and customer, there is a negotiation process using a Service Negotiation and Acquisition Protocol (SNAP) and depending on different scenarios. In the first scenario, all the users ask for the agreement using the agreement state transitions (where the coordinator plays the role of the service's broker). The two other scenarios are:

- Community Scheduler, which is an entity that acts as an intermediary between the community and its resources, and
- File Transfer; it restricts the activity of sending and receiving requests from/to the user (e.g. transferring a file) with a deadline time [10].

These two Scenarios are used to negotiate with the E-Learning services.

Figure (6) shows the negotiation process using the community scheduler for the process of uploading the material book, which is done by the instructor for the student where:

- Task Service Level Agreements **(TSLAs)** is the procedure that concerns with the task of the uploading process. So, TSLA is "Uploading Process".

- Resource Service Level Agreements **(RSLAs)** where it defines the available resources in the system. It depends on the procedure of the TSLA. So, the RSLAs is "RSLA Resources".

- Binding Service Level Agreements **(BSLAs)**; it is the binding agreement to apply the procedure of a service. So, the BSLA is: "upload the resources".

The activities diagram for the proposed SLA for the E-Learning system consists of three entities; the instructor '*i*,' the student '*u*', and the scheduler 's', where both the student and the instructor perform the task by sending their activity, and the scheduler manages the system during the process.

Figure (6) shows the material activities that consist of two main processes, uploading the material by the instructor, and download it by the student.

According to Figure (6), each user sends activation to the community scheduler to perform the task whether from the instructor (*i*) or student (*u*), where:

TSLA: upload/download process.

RSLA: content Materials.

BSLA: upload/download resources.

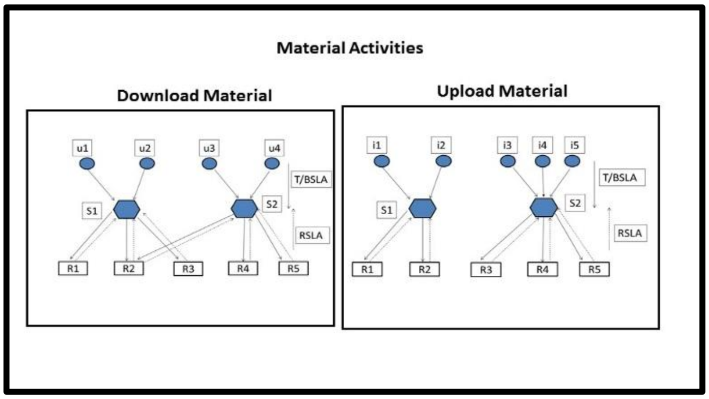


Fig. 6. Material Activities for E-Learning System

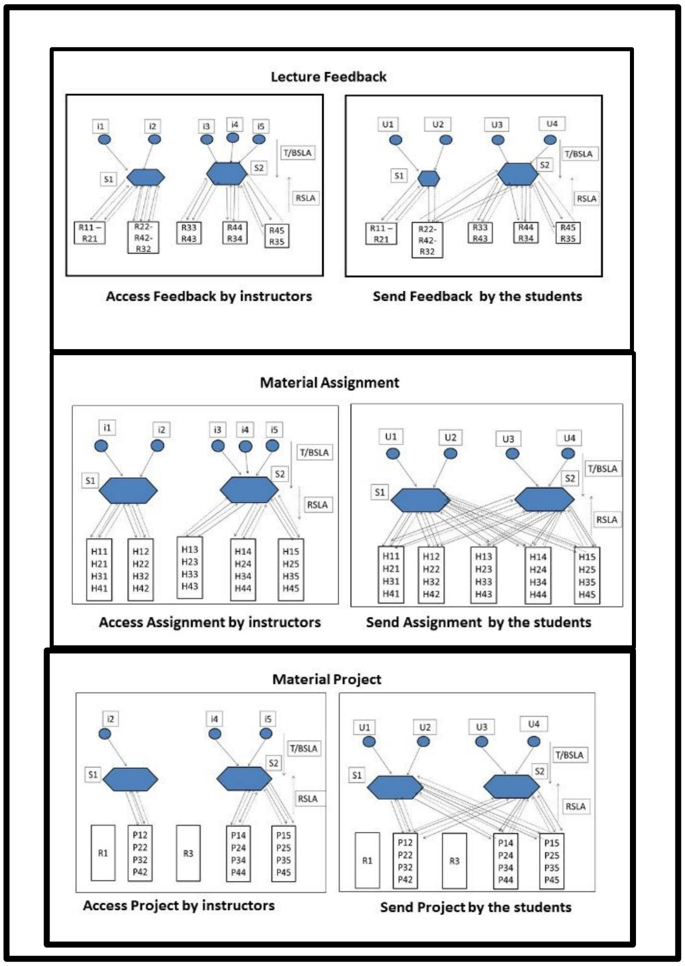


Fig. 7. Academic Activities For E-Learning System

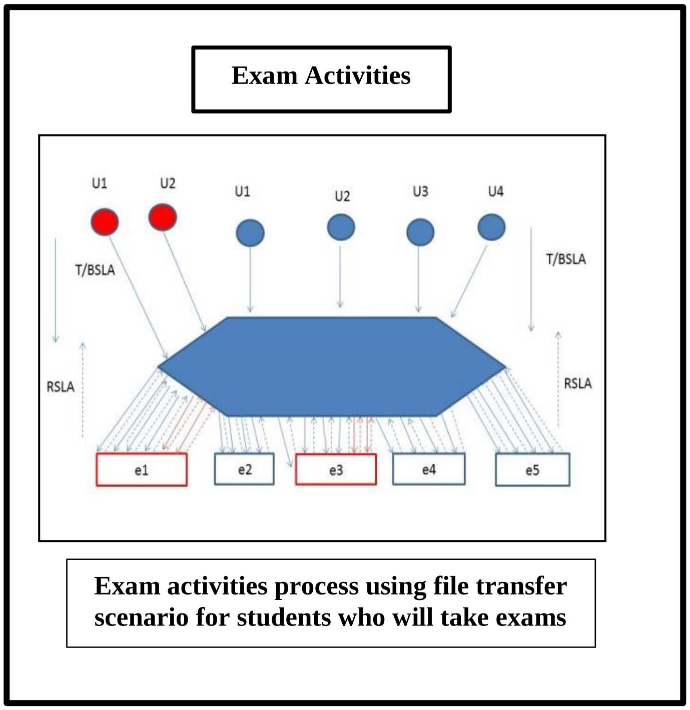

Exam Activities
U1
U2
U1
U2
U3
U4
T/BSLA
T/BSLA
RSLA
RSLA
e1
e2
e3
e4
e5
Exam activities process using file transfer scenario for students who will take exams

Fig. 8. Exam Activities for E-Learning System

Figure (7) represents the academic activities diagram for the proposed E-Learning system. It consists of three activities that are the lecture feedback, the material assignment, and the material project. For these activities, the student (*U*) sends an activation for the community scheduler (*S*) to perform the TSLA (i.e. the sending process), and also the instructor (*i*) sends an activation to access the RSLA (i.e. resources). Finally, the BSLA is performed by sending the lecture feedback, Material assignment, and Material project.

Figure (8) represents the exam's activity in the E-Learning system as an example. The user submits the job to the scheduler within a deadline time, and then the scheduler reserves a storage space on the destination resource to ensure there is enough space for the current activity before beginning the transfer. The file transfer scenario handle achieving this activity. Once the space is allocated, the scheduler reserves the suitable bandwidth from the network.

VI. SLA Framework Sequence Diagram

The sequence diagram is an interaction diagram shows how the service’s processes operate with one another with defining their orders [15]. Figure (9) describes the SLA frame work sequence diagram that explains the sequence of the services for all users (i.e. students, instructor, and the coordinator).

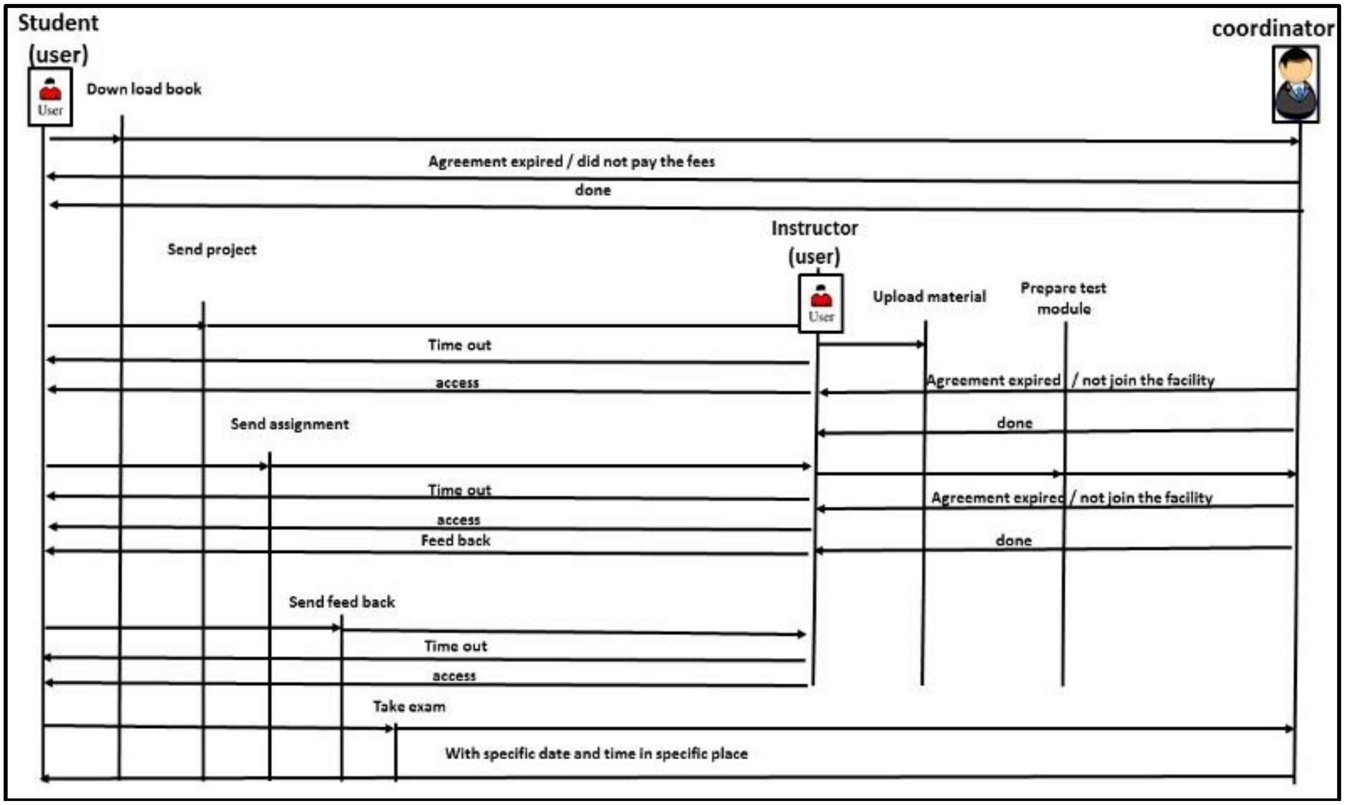


Fig. 9. Frame Work Sequence Diagram

VII. Conclusion and Future Work

An effective Service Level Agreement (SLA) is the way to guarantee the process of the service providing from the cloud provider to the cloud user. The more clear definition of the SLA parameters gives more flexibility, confidence, and reliability between the provider and the user. The main goal of this paper is to represent the service level agreement (SLA) for an E- learning system through defining the influential metrics of SLA for SaaS , IaaS, PaaS, and Storage as a

Service, and designing the framework, sequence diagram and architecture of the proposed system.

The E-Learning system implementation and SLA document (terms and conditions) will be considered as a future work.

REFERENCES

[1] Dillon, T., and W, C.," Cloud Computing: Issues and Challenges", IEEE International Conference on Advanced Information Networking and Applications, Vol.24, PP. 27-29, 2010.

[2] Jadeja, Y, and Modi, K" Cloud Computing-Concepts,Architecture and Challenges " International Conference on Computing, Electronics and Electrical Technologies (ICCEET), PP. 877-899, 2012.

[3] Prasanth, A.," Cloud Computing Services: A Survey", International Journal of Computer Applications, Vol. 46, No. 3, May 2012

[4] Pallis, G. "Cloud Computing the New Frontier of Internet Computing", IEEE Computer Society,PP.71-75,2010.

[5] Zhang,Q. and Boutaba ,R.," Cloud Computing: state-of-the-art and research challenges", Journal of Internet Services and Application, Vol.1,Issue1, PP. 7-18, 2010

[6] Alhamad, M., Dillon,T., and Chang,E.," Conceptual SLA Framework for Cloud Computing", IEEE International Conference on Digital Ecosystems and Technologies ,vol. 4th, PP. 606-609, TBD Dubai, United Arab Emirates 2010.

[7] POCATILU, P., Alecu, F.,and Vetrici, M.," Using Cloud Computing for E-learning Systems", Published in8th WSEAS international conference on Data networks, communications, computers,PP.54-59from November 7-9-2009

[8] Mathew,S., "implementation of Cloud Computing in education – A evolution", International Journal of Computer Theory and Engineering, Vol. 4, No. 3, pp. 473, June 2012.

[9] Armbrust, M., Fox,A., Griffith,R., Joseph,A., Katz,R., Konwinski,A., Lee,G., Patterson,D., Rabkin,A., Stoica,I.,and Zaharia,M., "Above the Clouds: A Berkeley View of Cloud Computing",http://www.moonther.com/cis492/abovetheclouds.pdf

[10] Czajkowski,k.,Foster,I., Kesselman,C., Sander,V.,and Tuecke,S., " SNAP: A Protocol for Negotiating Service Level Agreements and Coordinating Resource Management in Distributed Systems", Job Scheduling Strategies for Parallel Processing, Lecture Notes in Computer Science, Vol 2537, pp 153-183, 27 November 2002.

[11] Chenkang, W., Zhu,Y., and Pan,S., "The SLA Evaluation Model for Cloud Computing", International Conference on Computer, Networks and Communication Engineering, PP. 331-332, Beijing, China May 23-24, 2013.

[12] Patel,P ., A. Ranabahu , and A. Sheth , ” Service Level Agreement in Cloud Computing” kno.e.sis puplications http://knoesis.wright.edu/library/download/OOPS LA_cloud_wsla_v3.pdf (2009)

[13] Shrivastava, G.," E- Learning Improved Architecture for Clouds," Global Journal of Computer Science and Technology Cloud and Distributed, Vol. 13 Issue 2, PP. 33-35, 2013

[14] Rouse.M,http://searchcrm.techtarget.com/definition/entity-relationship-diagram, Posted in October 2014, last access on 22.8.2015 at 1:39am.

[15] Moreia, A., Araújo,A.,and Brito,I.," Crosscutting Quality Attributes for Requirements Engineering", Software Engineering and Kno wledge Engineering Conference(SEKE), ISBN:1-58113-556-4,PP 167-174, Ischia, Italy, 15-19 July 2002.

FRoTeMa: Fast and Robust Template Matching

Abdullah M. Moussa
Electrical Engineering Department
Faculty of Engineering,
Port-Said University,
Port-Said, Egypt

M. I. Habib
Electrical Engineering Department
Port-Said University, Egypt
Saudi Electronic University (SEU),
Saudi Arabia

Rawya Y. Rizk
Electrical Engineering Department
Faculty of Engineering,
Port-Said University,
Port-Said, Egypt

***Abstract*—Template matching is one of the most basic techniques in computer vision, where the algorithm should search for a template image T in an image to analyze I. This paper considers the rotation, scale, brightness and contrast invariant grayscale template matching problem. The proposed algorithm uses a sufficient condition for distinguishing between candidate matching positions and other positions that cannot provide a better degree of match with respect to the current best candidate. Such condition is used to significantly accelerate the search process by skipping unsuitable search locations without sacrificing exhaustive accuracy. Our proposed algorithm is compared with eight existing state-of-the-art techniques. Theoretical analysis and experiments on eight image datasets show that the proposed simple algorithm can maintain exhaustive accuracy while providing a significant speedup.**

Keywords*—template matching; pattern matching; brightness-contrast invariance; rotation invariance; scale invariance; sufficient condition

I. Introduction

Template matching is the task of seeking a given template in a given image. It is also known as pattern matching [1] and can be considered as one of the most basic operations in computer vision. Template matching is heavily used in signal, image and video processing. A lot of applications are based on template matching such as image denoising [2-4], motion estimation [5], and emotion recognition [6] to name a few. The literature on template matching contains a variety of algorithms. Based on the search accuracy, these algorithms can be broadly divided into two categories: approximate accuracy and exhaustive accuracy algorithms. The first category can achieve fast speedup at the cost of some loss of accuracy and often depends on one or more approximations, while exhaustive accuracy algorithms obtain fast speedup without losing accuracy. This category includes domain transformation techniques using FFT and bound based computation elimination algorithms in which inappropriate search locations are skipped from computations.

In general, regardless of the accuracy required, there are several ways to tackle the problem. Some techniques use the normalized cross-correlation (NCC) to handle brightness/contrast-invariant template matching. This can be made faster using bounded partial correlation [7, 8] or integral images [9]. Some other techniques depend on scale and rotation invariant key points [10], while some rely on previous segmentation/binarization of the image to analyze [11, 12], FFT [13], partial elimination [14], correlation transitivity [15], histogram [16, 17], circular and radial projections [18, 19], or auto-correlation [20]. However, in many cases, these algorithms cannot be used due to some image characteristics such as little grayscale variations and the resistance of the image to be binarized efficiently, or because they can't handle the rotation, scale, brightness and contrast (RSBC) invariant template matching problem, which is critical to several applications, or because they are unacceptably slow for many real world scenarios.

The tremendous amount of time-sensitive applications of template matching always pushes the need for faster techniques. So in this paper, we present a fast algorithm to solve RSBC invariant grayscale template matching problem. Our algorithm, named FRoTeMa (Fast and Robust Template Matching), is also rotation/scale-discriminating within a predefined level of accuracy, i.e., the method determines the scale and rotation angle of the template within the matching process. Although we argue on maintaining exhaustive accuracy only when both the image and the template are equivalent, experiments on eight image datasets show that the proposed simple algorithm can provide exhaustive-like search when some deteriorating conditions such as blurring or JPEG compression are applied on query images. We have compared our algorithm with eight state-of-the-art techniques. Experimental results reveal that, although its simplicity, the proposed algorithm also shows a significant speed-up.

The rest of the paper is organized as follows: Section II introduces the proposed template matching algorithm. Section III presents the complexity analysis of the new algorithm. In Section IV the experimental results are presented. Finally, conclusions are summarized in Section V.

II. The Proposed Frotema Algorithm

FRoTeMa depends on bound comparisons to achieve fast performance. Instead of checking the similarity between each pixel in the template with a corresponding pixel in each candidate block in the image to analyze, FRoTeMa uses sufficient condition to distinguish between candidate matching positions and other positions that cannot provide a better degree of match with respect to the current best-matching one.

To describe the condition used, Let $\boldsymbol{I}$ be the image to analyze, $\boldsymbol{T}$ the template, $\boldsymbol{I_b}$ a block in $\boldsymbol{I}$ with the same size of $\boldsymbol{T}$, $\boldsymbol{T_{sb}}$ a sub-template block of size $P \times Q$ pixels, T_{sbs} the summation of pixel intensities within $\boldsymbol{T_{sb}}$, $\boldsymbol{I_{sb}}$ the corresponding sub-block in $\boldsymbol{I_b}$ and I_{sbs} the summation of pixel intensities within $\boldsymbol{I_{sb}}$. We consider $\boldsymbol{T}$ and $\boldsymbol{I_b}$ to be equivalent under brightness/contrast variation if there is a contrast correction factor α > 0 and brightness correction factor β such that:

$$\boldsymbol{T} \;= \alpha\, \boldsymbol{I_b} \;+\; \beta\, \mathbf{1} \qquad (1)$$

where $\mathbf{1}$ is the matrix of ones [18]. Similarly, we have:

$$\boldsymbol{T_{sb}} = \alpha\, \boldsymbol{I_{sb}} \;+\; \beta\, \mathbf{1} \qquad (2)$$

Using (2) we conclude:

$$T_{sbs} = \alpha\, I_{sbs} + \beta\, PQ \qquad (3)$$

If we segment $\boldsymbol{T}$ into smaller blocks similar to $\boldsymbol{T_{sb}}$, and form a new vector $\boldsymbol{T_v}$ to store summation of pixel intensities of each one of these blocks and then do the same for $\boldsymbol{I_b}$ to form a similar vector $\boldsymbol{I_v}$, we can use (3) to write:

$$\boldsymbol{T_v} \;= \alpha\, \boldsymbol{I_v} \;+\; \beta\, PQ\mathbf{1} \qquad (4)$$

Eq. (4) represents the relation between $\boldsymbol{T_v}$ and $\boldsymbol{I_v}$ which is similar to the relation between $\boldsymbol{T}$ and $\boldsymbol{I_b}$ in Eq. (1). For a correct match block in $\boldsymbol{I}$, and because α, β, P and Q are constants, the normalized versions of $\boldsymbol{T_v}$ and $\boldsymbol{I_v}$ to zero mean and unit magnitude (denoted as $\boldsymbol{T_{vnor}}$ and $\boldsymbol{I_{vnor}}$ respectively) will be equivalent. Using a threshold ts to establish a minimum degree of similarity required to accept a match, we can write:

$$\boldsymbol{T_{vnor}} - \boldsymbol{I_{vnor}} \le ts\mathbf{1} \qquad (5)$$

Inequality (5) provides the condition used in the algorithm. That is, for each block in $\boldsymbol{I}$, if the condition is satisfied, this means that the block is a candidate to be a correct match, while if the condition is not reached; the block will be skipped without a need for further consideration. Using such condition not only can accelerate the search for the correct match, but also the usage of vector normalization will make the condition able to handle the possible variation in brightness/contrast between the template and candidate blocks.

A. Template Manipulation

Many template matching algorithms have to scan the image to analyze several times to be able to support rotation-scale invariance and this procedure may take much time to compute. FRoTeMa, Instead of doing so, relies on the template to handle the task. First, a prespecified range of scales and orientations should be set. The orientation and scale of the original template are changed according to each combination of angles and scales within the predefined ranges.

We have used the OpenCV [21] library for such rotation and scale operations. Then, a squared area in the middle of the resulting template is chosen and split into a grid of equally-spaced non-overlapping squared blocks that have a predefined side length. $Ƞ_b$ is used as one of the algorithm parameters to specify such side length. The grid layout is utilized to make use of (5) while making the implementation more simple. This grid can be split into different number of blocks. If the template is exactly the same as a part of the input image, only one block is sufficient to use (5), while under brightness/contrast variation, we have to use more blocks to be able to apply vector normalization in order to remove the brightness/contrast difference. In our experiments, we have used a grid of nine blocks. The centered block in such grid is called as the head block (Fig. 1). The summation of pixel intensities within each block of the grid is calculated, and then these new nine values are normalized to zero mean and unit magnitude. After this procedure ends, a vector of nine normalized values for each angle-scale combination should be calculated. Notice that the $Ƞ_b$ value must be the same for all calculated vectors. These vectors will be stored in a new vector $\boldsymbol{Tm_v}$. At the end of this computation, the resulting $\boldsymbol{Tm_v}$ vector is sorted with respect to the calculated values of head blocks.

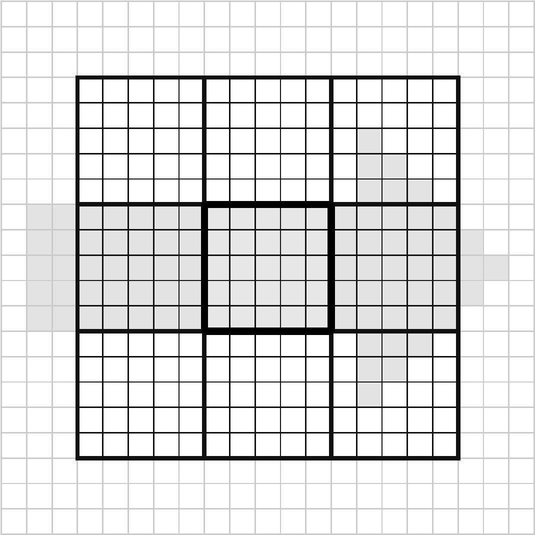


Fig. 1. A sample of 21×21 pixels template. FRoTeMa uses 9 blocks similar to big ones that appear in the middle. In this example, Ƞ_b is 5 pixels and head block is the one in the middle with a thicker edge

B. Image Scan

After computing $\boldsymbol{Tm_v}$, the image to analyze $\boldsymbol{I}$ is scanned sequentially in search of the template. Every pixel in $\boldsymbol{I}$ is checked only once by specifying nine squared blocks grid around it, calculating the summation of pixel intensities in such blocks and normalizing the resulting values to zero mean and unit magnitude. A new vector $\boldsymbol{P_v}$ is formed to always store the current pixel values. Similarly, the centered block of each pixel grid will be called as the head one. Notice that $Ƞ_b$ value should be fixed for the template and $\boldsymbol{I}$ grids. To accelerate the process, the grid of the left-most pixel in each row of $\boldsymbol{I}$ is computed. Then this grid is used to calculate its neighbor pixel's grid by an overlapping process using a sliding window and such procedure is continued with the same criterion until reaching the right-most pixel.

In the perfect case, when the change between the template and the matching block (if any) is only in brightness, contrast and/or scale-orientation within the prespecified ranges, the $\boldsymbol{P_v}$ vector at one or more pixels will be identical with at least one of the vectors of $\boldsymbol{Tm_v}$. In such case, only pixels which have a $\boldsymbol{P_v}$ vector that is identical to one or more $\boldsymbol{Tm_v}$ vectors will be interesting for further investigation, while in the common case (as seen in many real world scenarios), when some deteriorating conditions such as blurring or JPEG compression are applied on query images, the $\boldsymbol{P_v}$ vector may not be equivalent to any $\boldsymbol{Tm_v}$ vector. So, a thresholding technique is used in this case to handle such possible gap. Two thresholds t_h and t_o are used. t_h sets the limit for the absolute difference between the current pixel's head block value and $\boldsymbol{Tm_v}$ head blocks' values. If one or more $\boldsymbol{Tm_v}$ vectors pass such condition, the remaining eight values in $\boldsymbol{P_v}$ are checked against the corresponding eight values in the promising vector(s). t_o is used in such case to set the limit for the absolute difference between these values. Head blocks are handled uniquely as they should be the least distorted ones if the template is scaled or rotated with a value that is not covered in the prespecified ranges. While we can merge t_h and t_o values, t_h condition can prune the majority of the $\boldsymbol{Tm_v}$ vectors without a need for further processing. We call the overall of promising $\boldsymbol{Tm_v}$ vectors at all candidate pixels as candidate states. The amount of such states is important as it has a direct relation with speed of computation.

C. Candidate Pixels Handling

Each candidate pixel is in the center of a potential match block $\boldsymbol{b}$ in $\boldsymbol{I}$, and all of them have, by definition, one or more promising $\boldsymbol{Tm_v}$ vectors to be checked. When the current pixel is considered as a candidate one, the following is done with each one of its promising vectors: a new version $\mathbf{adj_t}$ of the original template is formed using the angle and scale that are associated with the vector. Before checking the matching between $\mathbf{adj_t}$ and $\boldsymbol{b}$, the possible variation in brightness and/or contrast between them should be handled. Several techniques may be used to handle this variation such as correlation coefficient and normalization. These techniques are efficient but they have expensive computational complexity. To avoid such drawback, another procedure is used based on (1) to estimate a contrast correction factor α_s and a brightness correction factor β_s. If such estimation relied on a few pixels, it will be sensitive to possible noise. So, a squared area $\boldsymbol{t_a}$ at the center of $\mathbf{adj_t}$ is chosen and intensity of $\boldsymbol{t_a}$ pixels is calculated after subtracting the mean value of them from each pixel to form $\boldsymbol{t_a}$\`. A similar process is done for a corresponding squared area $\boldsymbol{I_a}$ in $\boldsymbol{b}$ to form $\boldsymbol{I_a}$\`(We found experimentally that $\boldsymbol{t_a}$ and $\boldsymbol{I_a}$ of 11×11 pixels is far enough). This is made to cancel the possible variation in brightness between $\boldsymbol{t_a}$ and $\boldsymbol{I_a}$. To estimate α_s, a new matrix $\boldsymbol{C_m}$ is calculated as = $\boldsymbol{t_a}$\` / $\boldsymbol{I_a}$\`, and α_s is considered to be the median value of $\boldsymbol{C_m}$. α_s can be used to estimate β_s by using another matrix $\boldsymbol{B_m} = \boldsymbol{t_a}$ - ($\alpha_s \boldsymbol{I_a}$). The median value of $\boldsymbol{B_m}$ can be considered as an estimation of β_s.

To check whether $\boldsymbol{b}$ is a correct match block, the sum of absolute difference (SAD) is used between $\mathbf{adj_t}$ and $\boldsymbol{b}$ using the following formula:

$$SAD = \Sigma \,|\, (\alpha_s \boldsymbol{b} + \beta_s) - \mathbf{adj_t} \,| \qquad (6)$$

The SAD value is calculated within the area of the largest circle that can be fit in the middle of $\mathbf{adj_t}$ as such circled area will always be within the template if it is rotated. The ratio between the SAD value and the sum of pixel intensities within the mentioned circled area is calculated for each candidate state in search of the minimum ratio. If the minimum ratio is below some threshold t_s, the template is considered to be detected at the current pixel. The algorithm has been applied successfully on rescaled versions of both of the image to analyze and the template to be a one fourth of their original sizes. i.e., a one level of pyramidal reduction has been used. We didn't have to apply the algorithm on the original scale to get accurate results except in a few cases.

III. Complexity Analysis

Let N be the number of pixels in the image to analyze $\boldsymbol{I}$, M the number of pixels in the template image, C the number of candidate states. G the number of angles and S the number of scales. Practically, we can consider that G and S have a complexity of O($\sqrt{M}$) and η_b is, by definition of O($\sqrt{M}$). As N is usually much larger than M, G, S and η_b, all operations that do not depend on N or C are neglected. At the image scan step, the operations used to prepare each nine-blocks grid with the help of the sliding window have a complexity of O($N\eta_b$) or O($N\sqrt{M}$), while the operations used to calculate $\boldsymbol{P_v}$ have a complexity of O(N). The computation used to match $\boldsymbol{P_v}$ with $\boldsymbol{Tm_v}$ is of O($N \log GS$) or O($N \log M$) as we depend on binary search technique when searching for promising vectors in $\boldsymbol{Tm_v}$. Also, the algorithm uses O(MC) computation to handle the candidate pixels. Thus, the overall complexity of the algorithm is O($N\sqrt{M} + MC$). As will be clarified in the experiments, the proposed algorithm achieves a great reduction in C from millions or hundreds of thousands to just hundreds or tens. Thus, the overall complexity can be reduced to O($N\sqrt{M}$) as opposed to, for example, O(NM) in [19].

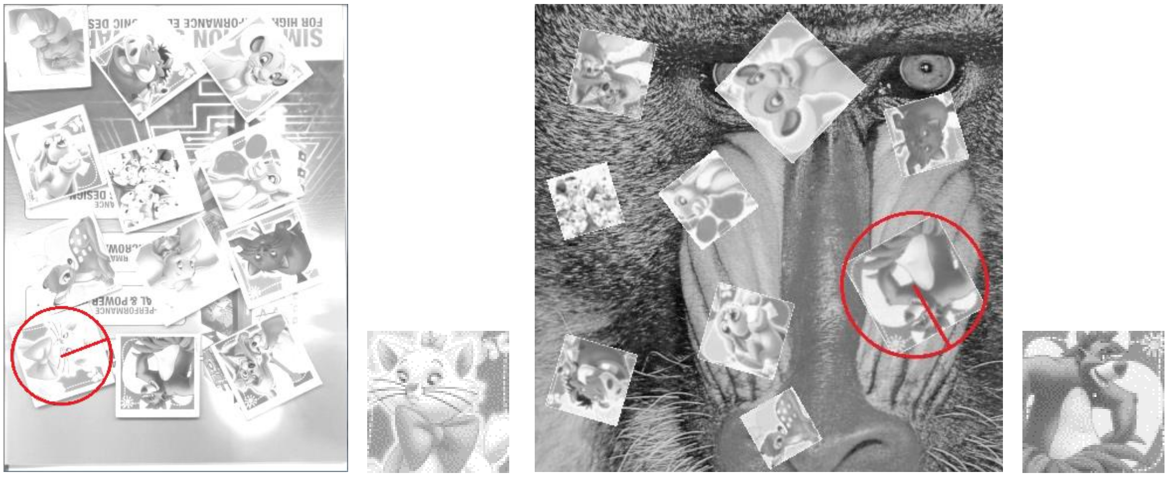
IV. Experimental Results and Discussion

A. Experiments

To check FRoTeMa performance, several experiments have been executed using eight datasets. All experiments have been run on a Core i-5 (2.3-GHz PC) with 4 GB of RAM. To check the proposed algorithm robustness to simultaneous variations in rotation, scale, brightness and contrast, a dataset (K1) provided by Kim [22] has been used. It consists of six images and twelve templates. Each one of the six images includes all of the templates. This dataset tests only the rotation invariance. So, for each query, the scale, brightness and contrast of the queried image have been changed randomly within the ranges of 0.7 to 1.4, -25 to +25 and 0.85 to 1.15 of the original values respectively and each randomly changed query has been tested twice.

We have compared FRoTeMa with eight state-of-the-art algorithms ([8, 13-15, 18-20]). Experiments with ZEBC [8], FFT [13], PCE [14], TEA [15], OptA [20] and EGS [20] are performed on satellite image dataset (SI) of a low population density seaport (962x622 pixels) and aerial image dataset (AI) of a densely populated area (1549x2389 pixels) [20] using implementations provided by [23]. For each of the two datasets, templates of 10 different sizes are used (ranges from 31x31 to 121x121 for SI and 39x39 to 129x129 for AI). In each size 10 templates have been tested. The templates are obtained from a different view point at a different time. Therefore both datasets contain projective distortions as well as illumination variations.

Comparison with Forapro [18] and Ciratefi [19] was against five datasets. To test and compare robustness to rotation and scale variations, we have used another dataset (K2) provided by H. Kim [22]. It consists of eight images (400×400 pixels) and nine templates (71×71 pixels). Each one of the eight images includes all of the templates. Fig. 2 shows some examples of the algorithm output against K1 and K2 datasets.



(a) (b)

Fig. 2. (a) An example of FRoTeMa output when queried using the illustrated template from K1. (b) Another example of FRoTeMa output when queried using the illustrated template from K2 (note the discriminated angles and scales)

We have also used four datasets provided by K. Mikolajczyk [24] to check how robust our algorithm is to variations in brightness-contrast (Leuven dataset), focus blur (Bikes and Trees datasets) and JPEG compression (UBC dataset). Each one of these four sets consists of 6 images. The amount of distortion increases as the image number increases from 1 to 6. We have extracted 20 random non-overlapped templates (89×89 pixels) from the first image of each one of the four datasets, and in each set, we have tested FRoTeMa along with Forapro [18] and Ciratefi [19] against all of the 6 images using all of the 20 templates. So, we have evaluated 360 queries for each set. Fig. 3 illustrates some examples of FRoTeMa output against the tested Mikolajczyk datasets. When testing FRoTeMa against K1 and K2, we have checked 36 angles and 6 scales due to datasets challenge, while it was more appropriate when querying the other datasets to check one scale and orientation as they almost don't change. Parameters of the tested algorithms have been used as recommended by [18], [19] and [23]. Table I summarizes the total execution time on AI, SI and K2 datasets of FRoTeMa and the mentioned eight known algorithms.

Table I shows how FRoTeMa can provide an exceptional speed-up over all of the tested existing algorithms. The overall numbers of possible matches when considering the number of pixels, scales and orientations checked for K2, Leuven, Bikes, Trees and UBC datasets are 34560000, 540000, 700000, 700000 and 512000 respectively, while the average number of candidate states tested for each single query when testing FRoTeMa against such datasets was just 432.1, 42.3, 50.6, 283.7 and 31.6 states respectively. This justifies why FRoTeMa provides such great speed which makes a much difference regarding time-sensitive applications.

For K1, K2, Leuven, Bikes, Trees and UBC datasets, the output of the algorithms was considered correct if the overlap error is less than 15%. This is a very strict condition if we realize that Mikolajczyk et al. [25] consider that regions with up to 50% overlap error can still be considered matched successfully using robust descriptors.

TABLE I. TOTAL EXECUTION TIME (SEC) AND SPEED-UP RATIO ON AI, SI AND K2 DATASETS. FROTEMA IS COMPARED WITH EIGHT EXISTING ALGORITHMS (SUR REFERS TO SPEED-UP RATIO)

	FRoTeMa	TEA	OptA	EGS	PCE	FFT	ZEBC
AI	57	237	267	408	668	1675	1680
SUR	1	4.1	4.6	7.1	11.7	29.3	29.4
SI	13	49	58	64	124	153	239
SUR	1	3.7	4.4	4.9	9.5	11.7	18.3
		FRoTeMa		Forapro		Ciratefi	
K2		21		87		282	
SUR		1		4.1		13.4	

For AI and SI datasets, the ground truth is not available, so a match was considered correct if it was within (±4; ±4) pixels of the output location of the OptA [20] algorithm that claims an exhaustive-like accuracy. The performance of algorithms is given in terms of recall: TP/(TP+FN), where TP is True Positive and FN is False Negative. In Table II, performance of FRoTeMa against Forapro [18] and Ciratefi [19] in the tested datasets is illustrated.

TABLE II. PERCENT PERFORMANCE OF THE PROPOSED ALGORITHM AGAINST FORAPRO [18] AND CIRATEFI [19] IN THE TESTED DATASETS. * MEANS THAT THE EXPERIMENT WAS NOT DONE AND NOQ REFERS TO NUMBER OF QUERIES

Dataset	NoQ	FRoTeMa	Forapro	Ciratefi
AI (illumination)	100	100	*	*
SI (illumination)	100	100	*	*
K1 (RSBC)	144	100	*	*
Leuven (camera aperture)	120	100	100	100
Bikes(focus blur)	120	100	100	100
Trees(focus blur)	120	100	97.5	100
UBC (JPEG compression)	120	100	97.5	100
K2 (rotation/scale)	72	100	100	100

Table II shows that besides its superiority in speed, FRoTeMa exhibited 100% accuracy.

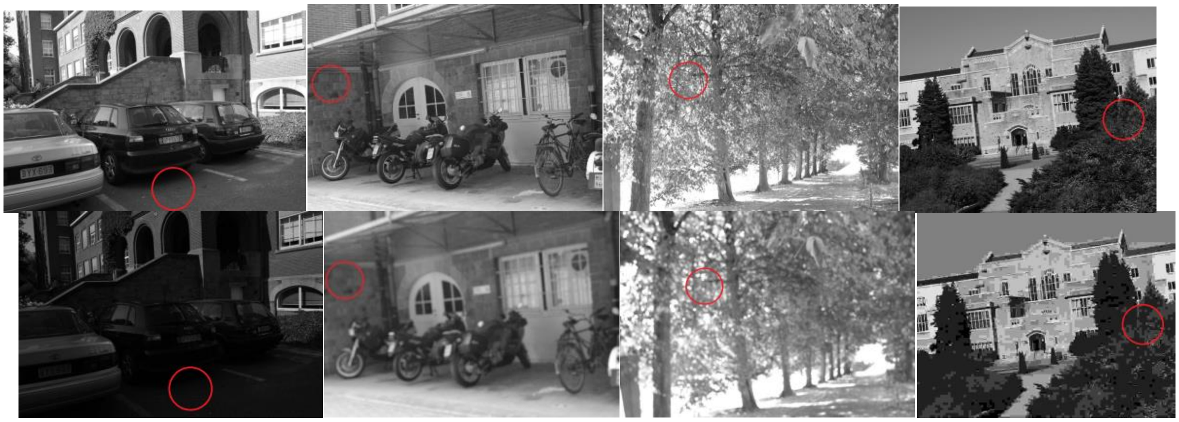


Fig. 3. Some examples of FRoTeMa output when tested against Mikolajczyk datasets [24]. For each couple of images, the upper one is the first image in the dataset and the lower one is the sixth image in the dataset when queried against the same template

B. Parameters

FRoTeMa sensitivity to different choices of η_b, t_h, t_o and t_s parameters has been tested. We have examined each one of the parameters while the others are fixed when searching for a template in one image of K2. We have checked 36 angles and 6 scales (from 0.7 to 1.4 of the original template size). In each of the following tables, the fixed parameters can be found in the first line.

TABLE III. SENSITIVITY TO η_b

$t_h = 0.03$, $t_o = 0.06$, $t_s = 0.09$		
η_b	Result (H refers to Hit)	No. of candidate states
3	H	1428
5	H	1997
7	H	2120
9	H	2280
11	H	2040
13	H	2232
15	H	2371
17	H	2449
19	H	2365
21	H	2034
23	H	2088

As we can see in Table III, changing η_b value almost has no effect of the result because no error was detected when varying its value. Also the resulting numbers of candidate states under η_b variation are comparable.

TABLE IV. SENSITIVITY TO t_h

$\eta_b = 9$, $t_o = 0.04$, $t_s = 0.04$		
t_h	Result (H refers to Hit)	No. of candidate states
0.005	Template Not Found	--
0.007	H	57
0.05	H	370
0.15	H	498
0.2	H	506
0.4	H	506
0.8	H	506
0.99	H	506

As of η_b we can see in Table IV that t_h doesn't affect the correctness of the results after some small value (0.007). Also we can note that the numbers of candidate states are comparable within a wide range of t_h values.

TABLE V. SENSITIVITY TO t_o

$\eta_b = 9$, $t_h = 0.007$, $t_s = 0.04$		
t_o	Result (H refers to Hit)	No. of candidate states
0.03	Template Not Found	--
0.04	H	57
0.06	H	555
0.08	H	2939
0.1	H	10536
0.12	H	30594
0.2	H	385920
0.4	H	1707086
0.6	H	2192439
0.8	H	2291491
0.99	H	2295690

Table V shows that t_o parameter has a broad range of values that will not affect the correctness (above 0.04). However, it has a big influence on the number of candidate states. So, large values of t_o may make the algorithm slower.

TABLE VI. SENSITIVITY TO t_s

$t_h = 0.007$, $t_o = 0.04$, $\eta_b = 9$		
t_s	Result (H refers to Hit)	No. of candidate states
0.03	Template Not Found	--
0.04	H	57
0.15	H	57
0.2	H	57
0.4	H	57
0.6	H	57
0.8	H	57
0.99	H	57

Table VI shows that after some value (around 0.04 in this test), the t_s value has no influence on the result. It also doesn't affect the number of candidate states. So, the choice of this value is not critical in terms of its effect on correctness or speed within a wide range of values.

V. CONCLUSIONS

Template matching plays a vital role in image processing and computer vision. It is heavily used in many applications in several diverse areas. In this Paper, We presented FRoTeMa, a new rotation, scale, brightness and contrast invariant template matching algorithm. The proposed method is also rotation/scale-discriminating within a predefined level of accuracy. Algorithm analysis and experimental results using eight datasets and comparisons with eight known methods show that FRoTeMa can provide an exceptional speed which is more suitable for time-sensitive applications while maintaining exhaustive-like search. Future work is aimed at checking whether the proposed method will be able to handle the color template matching problem and deal with color constancy. Also, it will be interesting to investigate the performance of the proposed algorithm after converting it to a parallel one.

REFERENCES

[1] Ouyang W, Tombari F, Mattoccia S, Stefano L, Cham W-K. Performance evaluation of full search equivalent pattern matching algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 2012; vol. 34: 127-143.

[2] Buades A, Coll B, Morel J-M. A Non-Local Algorithm for Image Denoising. IEEE Int. Conf. Comp. Vis. Pattern Recog. 2005; vol. 2: 60-65.

[3] Dabov K, Foi A, Katkovnik V, Egiazarian K. Image Denoising by Sparse 3D Transform-Domain Collaborative Filtering. IEEE Trans. Image Process. 2007; vol. 16: 2080-2095.

[4] Zhang R, Ouyang W, Cham WK. Image Deblocking Using Dual Adaptive Fir Wiener Filter in the DCT Transform Domain. in: Proceedings of IEEE Int'l Conf. Acoustics, Speech, and Signal Processing. 2009; 1181-1184.

[5] Shi X, Diwanji T, Mooney KE, Lin J, Feigenberg S, D'Souza WD, Mistry NN. Evaluation of Template Matching for Tumor Motion Management with Cine-MR Images in Lung Cancer Patients. Medical Physics. 2014; vol. 41.

[6] Mostafa E, Farag A, Shalaby A, Ali A, Gault T, Mahmoud A. Long Term Facial Parts Tracking in Thermal Imaging for Uncooperative Emotion Recognition. Sixth Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS). 2013; 1-6.

[7] Di Stefano L, Mattoccia S, Tombari F. ZNCC-based template matching using bounded partial correlation. Pattern Recognition Letters. 2005; vol. 26: 2129-2134.

[8] Mattoccia S, Tombari F, Di Stefano L. Reliable rejection of mismatching candidates for efficient ZNCC template matching. in ICIP. IEEE. 2008; 849–852.

[9] Lewis JP. Fast template matching. Proc. Vision Interface. 1995; 120-123.

[10] Lowe DG. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 2004; vol. 60: 91-110.

[11] Torres-Méndez LA, Ruiz-Suárez JC, Sucar LE, Gómez G. Translation, rotation, and scale-invariant object recognition. IEEE Trans. Syst., Man, Cybern. 2000; vol. 30: 125-130.

[12] Kim WY, Yuan P. A practical pattern recognition system for translation, scale and rotation invariance. in: Proceedings of Int. Conf. Comput. Vis. Pattern Recognit. 1994; 391-396.

[13] William P, Saul T, William V, Brian F. Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge University Press; 2007.

[14] Mahmood A, Khan S. Correlation-coefficient-based fast template matching through partial elimination. IEEE Trans on Image Process. 2012; vol. 21, no. 4: 2099 –2108.

[15] Mahmood A, Khan S. Exploiting transitivity of correlation for fast template matching. IEEE Trans on Image Process. 2010; vol. 19, no. 8: 2190 –2200.

[16] Ullah F, Kaneko S. Using orientation codes for rotation-invariant template matching. Pattern Recognition. 2004; vol. 37: 201-209.

[17] Yoo J, Hwang SS, Kim SD, Ki MS, Cha J. Scale-invariant template matching using histogram of dominant gradients. Pattern Recognition. 2014; vol. 47: 3006-3018.

[18] Kim HY. Rotation-Discriminating Template Matching Based on Fourier Coefficients of Radial Projections with Robustness to Scaling and Partial Occlusion. Pattern Recognition. 2010; vol. 43: 859-872.

[19] Kim HY, Araújo SA. Grayscale template-matching invariant to rotation, scale, translation, brightness and contrast. IEEE Pacific-Rim Symp. Image and Video Tech., Lecture Notes in Computer Science. 2007; 100-113

[20] Mahmood A, Mian A, Owens R. Optimizing Auto-correlation for Fast Target Search in Large Search Space. arXiv:1407.3535v2 [cs.CV] 2014.

[21] Open Computer Vision Library, [Internet]. [cited 2015 September] Available from: http://sourceforge.net/projects/opencvlibrary/ .

[22] Kim, HY. [Internet]. [cited 2015 September] Available from: http://www.lps.usp.br/hae/software/

[23] Mahmood, A. [Internet]. [cited 2015 September] Available from: http://www.csse.uwa.edu.au/~arifm/OptA.htm

[24] Visual Geometry Group, Robotics Research Group, Department of Engineering Science, University of Oxford. [Internet]. [cited 2015 September] Available from: http://www.robots.ox.ac.uk/~vgg/data/

[25] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers (IEEE), 2005, 27 (10),pp.1615-1630.

A Bayesian Approach to Service Selection for Secondary Users in Cognitive Radio Networks

Elaheh Homayounvala
Assistant Professor
Cyberspace Research Institute
Shahid Beheshti University
Tehran, Iran

***Abstract*—In cognitive radio networks where secondary users (SUs) use the time-frequency gaps of primary users' (PUs) licensed spectrum opportunistically, the experienced throughput of SUs depend not only on the traffic load of the PUs but also on the PUs' service type. Each service has its own pattern of channel usage, and if the SUs know the dominant pattern of primary channel usage, then they can make a better decision on choosing which service is better to be used at a specific time to get the best advantage of the primary channel, in terms of higher achievable throughput. However, it is difficult to inform directly SUs of PUs' dominant used services in each area, for practical reasons. This paper proposes a learning mechanism embedded in SUs to sense the primary channel for a specific length of time. This algorithm recommends the SUs upon sensing a free primary channel, to choose the best service in order to get the best performance, in terms of maximum achieved throughput and the minimum experienced delay. The proposed learning mechanism is based on a Bayesian approach that can predict the performance of a requested service for a given SU. Simulation results show that this service selection method outperforms the blind opportunistic SU service selection, significantly.**

Keywords—Cognitive Radio; Service Selection; Bayesian Networks

I. Introduction

Efficient utilization of the scarce frequency spectrum resources has been the main reason to attract many researchers' interest to study dynamic spectrum access methods [1, 2]. Measurements have shown that the licensed parts of the spectrum in which conventional spectrum allocation methods are used by operators, are under-utilized [3, 4]. Cognitive radio systems then are introduced to use the available spectrum temporarily not in use by licensed users, through intelligent channel sensing. In these systems, secondary users (SUs) are allowed to utilize the spectrum, as far as their transmissions do not have any harmful impact on the primary users' (PUs) operation [5, 6]. Therefore, the effective detection of the PU's spectrum opportunities (holes), also called channel sensing or opportunity discovery, has a critical role in efficient usage of the valuable under-utilized parts of the spectrum and there are many research works focused on this topic [7].

The accurate sensing of the present status of the licensed channel is not the only challenge in cognitive radio networks (CRN). Assuming that the spectrum usage granularity of SUs is infinitive, then the effective throughput of SUs in a CRN depends only on the total duration of unused parts of the channel, which itself depends on the PUs traffic load. However, in practice, there is a minimum free channel time interval which is needed for different services of SUs to operate properly. This minimum duration depends on the structure of the secondary network as well as the service types and QoS requirements of SU services. In this case, the effective throughput of SUs also depends on the distribution of free parts of the licensed channel (Fig. 1). Ideally, in order to increase the achieved throughput of SUs, it is better to decrease this minimum required time interval by choosing proper physical layer parameters and utilizing less delay sensitive services. In practice, however, there is no control neither on secondary network structure, nor on its service requirements. Secondary network is a network just like the primary one, with the only difference that SUs do not have a licensed spectrum to work on. Therefore, if SUs have some knowledge about the distribution of the free channel time intervals, then they can make a better decision on choosing the right service to be used in a specific time. It is difficult to directly inform SUs of PUs' dominant used services and statistics of free channel time intervals in each area, for practical reasons.

This paper proposes a learning mechanism embedded in SUs to sense the primary channel for a specific length of time. Then, the algorithm recommends the SUs upon sensing a free primary channel, to choose the best service in order to get the best performance, in terms of achieved throughput. The proposed learning mechanism is based on a Bayesian approach that can predict the performance of a requested service for a given SU.

The rest of this paper is organized as follows. In section II, Bayesian networks and their application in CRN are briefly reviewed. Section III describes the system model as well as the proposed method to help SUs make a decision on choosing the best service at each specific time. Section IV represents the simulation results of the proposed service selection algorithm and finally, the paper is concluded in section V.

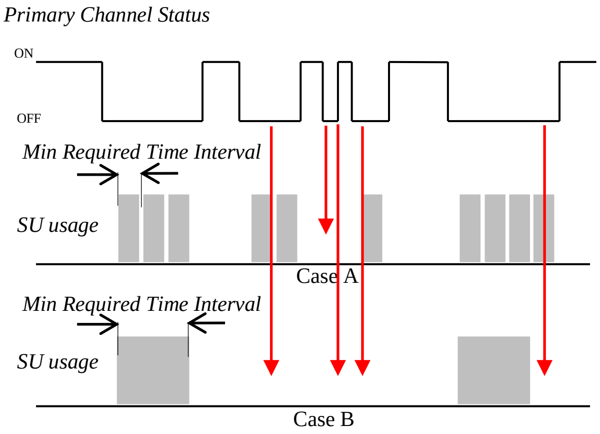


Fig. 1. Comparison of SU spectrum usage for different min required time interval values. Arrows show the "opportunities" of primary channel being off, but SUs cannot use the channel because of their min required time intervals

II. Bayesian Networks

Bayesian networks can deal with two problems of uncertainty and complexity. “The Bayesian network formalism was invented to allow efficient representation of, and rigorous reasoning with, uncertain knowledge” [8]. Bayesian networks are applied in many different applications and domains such as diagnosis, forecasting, automated vision, sensor fusion, manufacturing control and also mobile communications. Efficient monitoring, control and management of telecommunication networks [9], and network fault diagnosis, analysis and predictions [10, 11] are two main categories of applications of Bayesian network formalism in mobile communications. In addition to the above applications, we introduced a new application of Bayesian networks for modeling user preferences for radio access selection [12, 13] which was later extended [14, 15]. Bayesian networks have been applied in CRNs in localization [16], channel estimation [17], spectrum sensing [18] and channel selection [19] to name a few.

A Bayesian Network consists of two parts. The first part is the graphical representation, which is a directed acyclic graph. This is the qualitative part of the network. Secondly, conditional probability functions of each node in the graph, which form the quantitative part of a Bayesian network. In a Bayesian network, when node A is connected to node B with a directed arc from A to B, it usually means that node A causes node B. Node A is called the parent node and node B is called the child node. Conditional probability tables, in case of a discrete model, and conditional probability distributions in case of a continuous model, specify the probability of each child node conditioned on all possible combinations of values for all of its parent nodes.

The probabilities encoded by a Bayesian network can be learned from data. New information is combined with all previously known information using the Bayes’ theorem [20]. Thomas Bayes, an 18th century British mathematician, presented Bayes’ theorem, which is mathematically expressed as follows:

$$P(H\,|\,E,C)=\frac{P(H\,|\,C)P(E\,|\,H,C)}{P(E\,|\,C)} \tag{1}$$

Applying the Bayes’ theorem we can update our belief in hypothesis H given an additional evidence E and the background context C. P(H|E,C) represents the posterior probability of the hypothesis given the evidence. P(E|H) is the likelihood of the evidence given the hypothesis. P(H) represents the prior probability of the hypothesis and P(E) is the normalizing constant. Probabilistic inference is the most common task performed by the aid of Bayesian networks. Inference in Bayesian networks answers the questions of the probability of a variable based on given observations of other variables.

III. System Model

In this section the Bayesian network applied for proposed service selection algorithm for SUs in a CRN is explained. Bayesian networks have the ability to address problems of uncertainty. The uncertainty in this problem domain comes from two aspects. First there is uncertainty in service type used by PUs or requested by SUs. Secondly, call duration time and call arrivals for PUs are random. Therefore, the Bayesian view of the probability, that interprets probabilities as the “degree of belief” about events in the world, seems suitable for SUs' service selection. In such model, data is used to strengthen or weaken these degrees of belief.

The Bayesian network designed for this problem is illustrated in Fig 2. Without loss of generality, this paper defines the metric: *channel utilization* (CU) for each service in the secondary network, as the ratio of the time that channel is used for that service, to the time that channel is free for secondary usage. Therefore, this metric is independent of the primary traffic load and shows how efficient a secondary network with a specific service can use the primary spectrum holes. As it can be seen in this model, the CU of the channel depends on two parameters of PU service ratio (PUSR) and SU service type (SUST). PUSR_i is the ratio of the time that primary channel is occupied to service_i to the whole observation window. SUST represents the type of services that SUs can request.

Hence, the Bayesian inference in this model can answer the questions of the probability of the channel utilization being more than a given threshold for given service type *Sj* and a given PUSR. PUSR must be measured by sensing the channel for a specific window frame. Then, we refer to conditional probability table of CU node to see the value of the following term, for a given threshold:

$$\text{Predicted\_CU}_{\text{th},j} = \text{P}\ (\text{CU}\ |\ \text{PUSR} = \text{p}_i,\ \text{SUST} = \text{s}_j) \quad (2)$$

This is the probability of CU, predicted by our model to be used for the j^{th} service selection. In particular, if the value of this $\text{Predicted\_CU}_{th,i}$ is more than a threshold, then S_i is recommended to be used in secondary network, otherwise S_i is not recommended, which means that in this case, in order to have a CU more than the given threshold, this is not recommended to use S_i in secondary network. This test can be checked for different SUTS values, when there are more than two services in the secondary network.

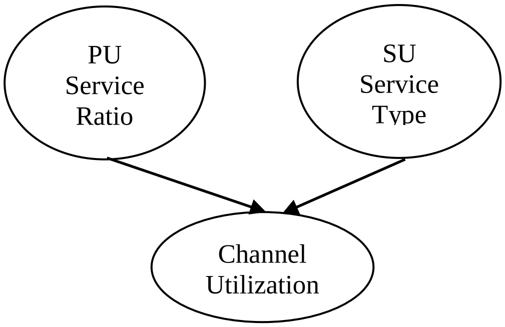


Fig. 2. Bayesian network for SUs' service selection in CRN

The smaller “mean call duration time” in the j^{th} service, and the more delay resistant the j^{th} service, the larger Predicted_CU$_{th,j}$. Generally speaking, in secondary networks, using delay tolerant services by short mean call durations has the lowest risk of blocking and yields to the largest channel utilization. However, when there are different types of services with different QoS requirements queued at SUs, it is important to choose the best service at each time in order to reach the largest throughput and the smallest delay for services. Therefore, when there are more than two services in the secondary network, service selection does not depend on Predicted_CU$_{th,j}$ only, but also on the j^{th} queue length as well.

IV. Simulation Results

In this section the simulation setup and its parameters are explained. A primary channel with the states of “1” and “0” is assumed to show the “busy” and “free” status, respectively. It is also assumed that the primary channel is used by PUs with two types of services; service1 whose call arrival times follow a Poisson distribution with $\lambda_1 = 10$ calls/min with exponentially distributed call duration times with a mean of $T_{1ave} = \Delta t$ sec, and service2 with $\lambda_2 = 0.5$ and $T_{2ave} = 50\Delta t$, where Δt is the time resolution of our time driven simulation. Service1 represents a typical short term service with a frequent call arrival, such as voice and some interactive gaming. Notice that the important point here is the traffic statistics rather than the QoS requirements of the service, as in a CRN, the PUs' services always have the priority over SUs ones, even though they are not delay sensitive, inherently. Service2 represents a typical long term service whose packets are transferred in bulks, such as file transfer or video downloading. These services take a more considerable continuous time interval of the channel, independent of if they are delay tolerable or not. Then, we have modeled a time window in which the channel is occupied by these two services, service1 and service2, for a given probability. Figure 3 shows primary channel status for a typical 50% mixed traffic case of these two services.

For secondary network, we assume the same two services as the primary network uses, for the sake of simplicity. We also set the min required time intervals of service1 and service2 to $T_{1min} = T_{1ave}$ and $T_{2min} = T_{2ave}$, respectively. We assume that SUs always have data packets of service1 and/or service2 in queue, ready to use the free primary channel. Table 1 shows the SU channel utilization (CU) for service2 for different combinations of service1 and service2 traffics in the primary channel.

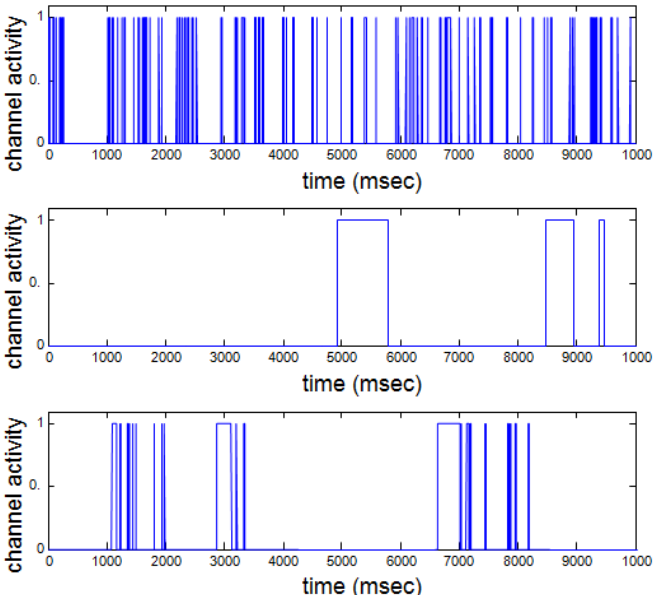


Fig. 3. A typical primary channel status: a) 100% service1 b) 100% service2 c) 50% mixed scenario

network. For service1, CU is 0.98 in all PUSR values. It can be seen from these conditional probability table that channel utilization for service1 SUs is always high, independent of service combination in the primary network. This is because $T_{1min} << T_{2min}$ in our simulation, which means that service1 can use most of the primary channel opportunities while service2 loses some of these opportunities because of its longer min required time interval. This loss increases as the ratio of service2 in primary channel traffic goes up.

In the last part of the simulation, two CRNs are compared. The first one uses a blind decision making about SUs services and the second CRN chooses the SUs service based on the Bayesian learning algorithm whose parameters are tuned according to the results of an observation time such as what have been shown in Fig. 3. The simulation results for an observation window, which contains at least 1000 channel status changes, show that the utilization of the proposed service selection algorithm outperforms the random service selection method for at least 20%.

TABLE I. Typical Values of Channel Utilization (CU) Using Service2 for Different Combinations of Service1 and Service2 Traffics in the Primary Network. Itr#i is the *ith* Iteration

PUS R	itr# 1	itr# 2	itr# 3	itr# 4	itr# 5	itr# 6	itr# 7	…	itr#10 0
80%	.82	.83	.79	.87	.86	.76	.84		.75
85%	.75	.84	.79	.81	.77	.73	.74		.80
90%	.51	.65	.73	.56	.71	.83	.64		.71
92%	.73	.67	.73	.72	.65	.63	.65		.88
95%	.43	.51	.36	.29	.82	.54	.44		.43
98%	.43	.25	.32	.33	.30	.19	.29		.28
99%	.19	.13	.16	.15	.10	.18	.16		.25
100%	.05	.05	.04	.04	.02	.02	.01		.04

V. Conclusion

This paper utilized the statistics of the primary channel occupancy to choose the secondary services in a way that the channel utilization of secondary users in a cognitive radio network increases. We have used a Bayesian network to model the channel utilization based on different possible services of primary and secondary networks. The simulations are performed for a simple case of two services, but they can be easily extended in future work to the case of multiple services and users in the primary network, to enhance the performance of the system in terms of dropping probability and experienced delay, as well.

References

[1] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohantly, ”Next generation/ dynamic spectrum access/cognitive radio wireless network:a survey,” *Elsevier Computer Networks*, vol. 50, pp. 2127-2159, Sept. 2006.

[2] A. Attar, M. S. A. Ghorashi, Sooriyabandara and A. H. Aghvami, ”Challenges of real-time secondary usage of spectrum”, *Computer Networks*, vol. 52, no. 4, pp. 816–830, 2008.

[3] M. McHenry, “Frequency agile spectrum access technologies,” *in Proc. FCC Workshop on Cognitive Radio*, May 2003.

[4] M. A. McHenry, “NSF spectrum occupancy measurements project summary,” *Shared Spectrum Company technical report*, Aug. 2005. Available at http://www.sharedspectrum.com/

[5] J. Mitola, “Cognitive radio for flexible mobile multimedia communications, ” *in Proc. IEEE Int. Workshop Mobile Multimedia Communications*, pp. 3-10, 1999.

[6] J. Mitola, “Cognitive radio: an integerated agent architecture for software defined radio, ” *Doctor of Technology*, Royal Inst. Technol. (KTH), Stockholm, Sweden, 2000.

[7] T. Yucek and H. Arslan, “A survey of spectrum sensing algorithms for cognitive radio applications, ” *IEEE Communications Surveys and Tutorials*, vol. 11, no. 1, pp. 116-130, First Quarter 2009.

[8] S. Russell and P. Norvig, *Artificial Intelligence a Modern Approach*, Prentice Hall series in Artificial Intelligence, 2003.

[9] A. Bashar, GP. Parr,S.I McClean, B.W. Scotney, D. Nauck, "Knowledge discovery using Bayesian network framework for intelligent telecommunication network management" *in Proc. Of Springer LNCS LNAI series, 4th Internaional Conference on Knowledge Science, Engineering anf Management (KSEM 2010)*, Belfast, UK, pp. 518-529, September 2010.

[10] R.M. Khanafer, B. Solana, B., J. Triola, R. Barco, L. Moltsen, Z. Altman, P. Lazaro, " Automated diagnosis for UMTS networks using Bayesian network approach", *IEEE Trans. Vehicular Technology*, vol. 57, no. 4, pp. 2451 – 2461, 2008.

[11] A. Sedano-Frade, J. González-Ordás, P. Arozarena-Llopis, S. García-Gómez and A. Carrera-Barroso, "Distributed Bayesian diagnosis for telecommunication networks", *Advances in Practical Applications of Agents and Multiagent Systems, AISC*, vol. 70, pp 231-240, 2010.

[12] E. Homayounvala, A.H. Aghvami, "User preference modelling for access selection in multiple radio access environments", *IEICE Transactions on Communications, Special Section on Software Defined Radio Technology and Its Applications*, vol. EE88-B, no. 11, Nov. 2005, pp. 4186-4193.

[13] E. Homayounvala, S. A. Ghorashi, A. H. Aghvami, “A Bayesian approach to modelling user preferences for reconfiguration”, *E2R Workshop*, Barcelona, Spain, September 2004.

[14] A. Koutsorodi, E. Adamopoulou, K. Demestichas, M. Theologou, “Service configuration and user profiling in 4G terminals”. *Wireless Personal Communications*, Vol 43, No 4, p.p.1303-1321, Dec. 2007

[15] K. Demestichas, A. Koutsorodi, E. Adamopoulou, M. Theologou, “Modelling user preferences and configuring services in B3G devices”, *Wireless Networks*, July 2007.

[16] S. Kandeepan, et. al. “Bayesian Tracking in Cooperative Localization for Cognitive Radio Networks, ” IEEE 69th Vehicular Technology Conference, VTC Spring Barcelona, pp. 1 – 5, 26-29 April 2009.

[17] P. Demestichas et. al, “Enhancing channel estimation in cognitive radio systems by means of Bayesian networks, ” Journal of Wireless Personal Communications, vol. 49, no. 1, April 2009.

[18] Zhenghao Zhang Husheng Li Depeng Yang Changxing Pei, “Space-time Bayesian compressed spectrum sensing for wideband cognitive radio networks, ” IEEE Symposium of New Frontiers in Dynamic Spectrum, pp. 1 – 11, Singapore , 6-9 April 2010.

[19] Guangxiang Yuan et. al., “Multi-User cooperation for channel selection in cognitive radio networks: a Bayesian approach, ” IEEE Globecom 2010.

[20] R. L. Winkler, “Introduction to Bayesian Inference and Decision”, *Series in Quantities Methods for Decision Making, Holt*, Rinehart and Winston Inc, 1972.

A Real-Time Face Motion Based Approach towards Modeling Socially Assistive Wireless Robot Control with Voice Recognition

Abhinaba Bhattacharjee[1]
Department of Electronics and Communication Engineering
Adamas Institute of Technology, Barasat
West Bengal University of Technology, Kolkata –700064
Kolkata, West Bengal, India

Partha Das[2]
Department of Computer Science and Engineering
Regent Education and Research Foundation, Barrackpore
West Bengal University of Technology, Kolkata –700064
Kolkata, West Bengal, India

Debasish Kundu[3]
Gobindapur Saphali Memorial
Polytechnic, Gobindapur.
Guskara, Burdwan, West Bengal,
India

Sudipta Ghosh[4]
Bengal Institute of Technology and
Management, Santiniketan,
Gopalnagar, Birbhum, West Bengal,
India

Sauvik Das Gupta[5]
School of Electrical and Computer
Engineering,
Oklahoma State University
Stillwater, OK, USA

Abstract—**The robotics domain has a couple of specific general design requirements which requires the close integration of planning, sensing, control and modeling and for sure the robot must take into account the interactions between itself, its task and its environment surrounding it. Thus considering the fundamental configurations, the main motive is to design a system with user-friendly interfaces that possess the ability to control embedded robotic systems by natural means. While earlier works have focused primarily on issues such as manipulation and navigation only, this proposal presents a conceptual and intuitive approach towards man-machine interaction in order to provide a secured live biometric logical authorization to the user access, while making an intelligent interaction with the control station to navigate advanced gesture controlled wireless Robotic prototypes or mobile surveillance systems along desired directions through required displacements. The intuitions are based on tracking real-time 3-Dimensional Face Motions using skin tone segmentation and maximum area considerations of segmented face-like blobs, Or directing the system with voice commands using real-time speech recognition. The system implementation requires designing a user interface to communicate between the Control station and prototypes wirelessly, either by accessing the internet over an encrypted Wi-Fi Protected Access (WPA) via a HTML web page for communicating with face motions or with the help of natural voice commands like "Trace 5 squares", "Trace 10 triangles", "Move 10 meters", etc. evaluated on an iRobot Create over Bluetooth connectivity using a Bluetooth Access Module (BAM). Such an implementation can prove to be highly effective for designing systems of elderly aid and maneuvering the physically challenged.**

Keywords—Face detection; Skin tone segmentation; Voice Commands; Speech Recognition; Wi-Fi Protected Access (WPA); Arduino Wi-Fi Shield; iRobot Create; Surveillance systems

I. Introduction

In today's age, the robotic industry has been developing many new trends to increase the efficiency, accessibility and accuracy of the systems in order to automate the processes involved in task completion. With the improvement of the advancing technology, humans have inherited a tendency of reducing physical and mechanical efforts to avoid repetitive jobs that are boring as well as stressful. Since long Robots have been built to reduce tedious human efforts and human errors. Though robots can be a substitute to manpower, they still need to be controlled by the humans, whether the robots be wired or wireless and thus needs to be handled by a controller device, both having pros and cons associated with them. The complexity of calculations and computations has made a breakthrough in modern technology, and human-computer interactions using natural resources has been made possible with user-friendly interfaces and machine learning algorithms. With these algorithms, system training is made effective as the system responds to simple human gestures or colors available in the environment with corresponding outputs and hence can prove efficient in controlling man-machine interactions. So beyond controlling the robotic system through physical or electronic devices, if the recent gesture control method is applied to embedded robotic systems then it provides a rich and intuitive form of interaction with the system which mainly involves Image Processing and Machine Learning algorithms for application development. Beyond this, it also requires some hardware and software interfacing with the system for gesture acquisition and corresponding control signal generation. Many attempts have been made in this field using motion sensing apparatus like that of an accelerometer with a combination of the gyroscope. These have been conventionally used in various systems of Human-Computer Interaction for sensing and tracking hand gestures externally to direct mechanical Robots to desired directions. However, acquisition of speech signal and emulation of various system navigation events using Digital Speech Processing has also been a method of Human System Interaction. But this method did not prove to be effective as it is very much prone to the environmental noise and may result in inefficient outcomes until and unless Speech recognition is taken into account.

In reference Arce [1] et al, made an intuitive effort by carrying extra circuitry attaching a number of accelerometers on the hand to develop a hand gesture recognition technique using ANN. Hand gesture recognition using image processing algorithms many a times involve the use of colored gloves. By tracking this color glove, different hand gestures can be interpreted as described by Luigi Lamberti1 and Francesco Camastra in their paper [2] where they have modeled a color classifier performed by Learning Vector Quantization. Kim [3] et al. developed a pattern recognizing algorithm that has been used to study the features of the hand. Some gesture recognition systems involve adaptive color segmentation [4], hand finding and labeling with blocking, morphological filtering, and then gestures are found by template matching. These processes do not provide dynamicity for the gesture inputs. While in another approach, gestures are recognized using Microsoft Xbox 360 Kinect(C)[5]. Kinect gathers the color and depth information using an RGB and Infra-Red lens respectively. There are many papers where training of hands using a large database of near about 5000-10000 positive and negative images is considered. But this hand gesture recognition technique doesn’t provide an identical real-time biometric logical authorization to access control of surveillance systems, which made us shift the Gesture acquisition domain from Hand tracking to real-time face tracking as well as the acquisition of definite voice commands using digital speech recognition algorithm.

In our approach we have developed an intelligent user interface which involves two distinct and intuitive forms of interaction – either by acquiring real-time three-dimensional face motions by real-time face tracking algorithm, or by definite voice commands using digital speech recognition algorithm. Both the processes provides a unique logical authorization of the user to access a control station.

The system implementation involves the design of an advanced User Interface to access a control station to control an embedded wireless robotic prototype wirelessly, either using a Bluetooth access module or an encrypted Wi-Fi Network using a Wi-Fi shield with the help of either face gesture acquisition or specific voice commands. We have tested our algorithms on two different embedded platforms. The Speech recognition algorithm is implemented on a ready-made platform of *iRobot Create* using a wide range Bluetooth Access Module, while the real-time face detection and tracking algorithm has been evaluated on a self-developed embedded prototype built on an open source AVR microcontroller-based platform *(ARDUINO)* with an Arduino Wi-Fi shield to connect to a protected Wi-Fi network and control its navigation over the web via an HTML Web page. The next Section describes the Interfaces involved in the system design with Section (III) elucidating the Hardware Platform of the system followed by Section (IV) describing the methodology to design the overall system implementation. After that comes the Experimental Evaluations and Results in Section (V) with finally the Conclusion and the Future Works of the paper.

II. Interfaces

Interfacing is an integral part of every embedded electro-mechanical systems controlled by either automation or Human Computer Interaction. Our work is manly composed of two fundamental interfaces namely *User Interface* and *Wireless Interface.*

A. User Interface

Non-invasive techniques for controlling are in high pace with the advancement of technology. Many works have already been done by Computer Vision experts that include Augmented Reality, Controlling PC-Mouse events through color gestures [6] or hand gestures that include selecting, opening, closing files. Both mouse and keyboard can be replaced by virtual keyboard and mouse described by Tsang, W.-W.M [7] which reduces hardware components of a PC.

In our approach the user interface involves the interaction between the User and the Control station through face gestures or voice commands. Here we have used a Personal Computer (PC) to serve the purpose of control station for accessing control over the robotic prototype.

In case of control through real-time face detection and tracking, the PC webcam is used to collect data in the form of images in a live video stream. This data is further processed and computed by a real-time face detection and tracking algorithm using image processing [8] and computer vision algorithms to provide the prototype the exact control signal of what the user wants to direct through various face gestures.

While in the case of control through definite voice commands, the inbuilt PC microphone is accessed to acquire the speech signals in the form of sinusoidal waveforms, process them using speech recognition algorithm and generate control signals to direct the prototype in a specific direction according to the delivered voice command.

B. Wireless Interface

This includes the wireless connectivity of the embedded robotic prototype with the control station either by Bluetooth connectivity or through internet over a protected Wi-Fi network. In our work we have used a Bluetooth Access Module to connect the control station i.e. the PC to the readymade robot of *iRobot Create* and control it through Voice commands, while we have used a wireless Wi-Fi network for wireless hardware to hardware interfacing between the control station and the prototype developed on Arduino based platform.

For the second case both the prototype and the control stations are connected to the same Wi-Fi network and control signals generated through tracked real-time face gestures are transferred in TCP/IP in the form of data packets via a HTML web page over a secured Wi-Fi Protected Access encryption mode.

III. Hardware Platform & System Requirements

The system design of our project work involves a synchronization of hardware and compatible software integration for its effective working with satisfying results. Tested on both a readymade platform as well as an open source platform, the system includes the following specifications.

A. Control Station

In our work we have used a Personal Computer (PC) for voice as well as real-time face gesture acquisition to work as our control station with the following minimum hardware requirements:-

i. Processor Clock speed should be at least 2 GHz (Intel Core2Duo or Higher versions)
ii. RAM (Primary Memory) should be not less than 2GB.
iii. The OS (Operating System) should be compatible to support Arduino IDE 1.0.2v and higher versions of MATLAB (Windows7 onwards, Linux, MAC).
iv. A Digital Megapixel Front Web camera is needed to detect and track Faces.
v. An inbuilt microphone to acquire voice commands for processing speech recognition.

Fig. 1. iRobot Create

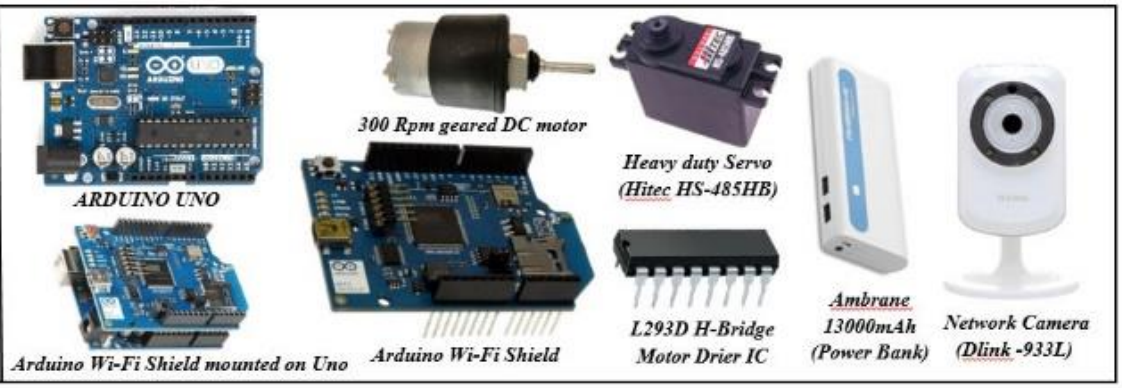


Fig. 2. The Hardware setup for the overall implementation of Arduino based Self-developed Robot

B. Prototype Platforms

The design implementation has been evaluated and simulated in two different microcontroller based platforms in order to check Mechanical Robustness, performance levels, flexibility and feasibility of the system.

- ❖ <u>*iROBOT – CREATE*</u>$^{®}$ *(Readymade Hobbyist Robot)*

This is a total set of robot development kit with a complete mechanical assembly developed by iRobot$^{®}$ based on the iRobot's Roomba Vacuum cleaning platform. It is rather known as a hobbyist robot with which one can develop new Robotic behaviours without worrying about the low level control. The hardware specifications include:-

i. *Cargo Bay* – which houses external electronic hardware like robotic arm or, sensors and actuators for external attachment.
ii. *DB-25 Port* – for providing serial communication to attached peripheral electronics.
iii. *7 Pin Mini DIN Connector* – A serial port through which sensor data can be read and motor commands can be issued using iRobot Roomba's Open Interface (ROI) protocol.
iv. An Omnidirectional IR Receiver – for obstacle detection or wall sensing.
v. Bluetooth Access Module (BAM) – This is a high range Bluetooth connectivity device manufactured by Element Direct. This hosts an individual client within an optimum range.

In addition to these features the platform accepts virtually all accessories designed for iRobot's second generation Roomba 400 Series domestic robots and can also be programmed with iRobot's own Command Module (a microcontroller with a USB connector and four DE-9 expansion ports).

The present day models replaced iRobot Roomba's Open Interface (ROI) and has provided Create's own Software interface that allows Create's behaviour manipulation through a series of commands including mode commands, actuator commands, demo commands, and sensor commands that is sent to Create's serial port by way of a PC connected to the DB-25 port using a BAM.

- ❖ <u>ARDUINO based</u> *(Self-Developed Robot)*

This is a Prototype built on an Open Source Microcontroller based platform (Arduino UNO) and especially designed to provide a facility of Navigational control through the Wi-Fi based web access using internet protocol.

The Hardware requirements involve:-

i. *ARDUINO Uno R3* – which primarily consists of Atmel Atmega 328p which is a 28 pin 8-bit AVR-RISC based microcontroller operating at a clock speed of 20 MHz.
ii. *Arduino Wi-Fi Shield (Compatible to Arduino Boards)* – Allows Arduino boards to connect to the internet

wirelessly through IEEE 802.11b/g/n system in package using both the TCP and UDP protocols.

iii. *L293D* – Motor Driver circuit, a 16 pin H-bridge IC to provide sufficient power output to the DC motors.

iv. *DC Motors (300RPM, 12V)* – a pair of geared DC motors providing a maximum of 300RPM at 12 Volts.

v. *Servo Motor (Hitec HS-485HB)* – a 3pole top ball bearing heavy duty high performance 180 degree servo motor providing a maximum torque of 6.4 Kg/cm at 6 Volts.

vi. *External Power Bank (Ambrane 13000mAh)* – To provide required power to all the circuitry operating at 5 Volts.

vii. *9V Batteries* – a set of 3 batteries in series is need to provide an uninterrupted 12 Volts potential difference to the pair of geared DC motors.

viii. *Network Camera (DLink DCS-933L)* – This is a Wi-Fi supported Day/Night surveillance camera with IR night vision which acquires live video footage of the prototype end and communicates it to the control station end remotely.

The primary motive for such a design is to develop an interactive embedded robot operated over the web through Wi-Fi, with a surveillance camera for live prototype feed, which can serve the purpose of various industrial applications as well as be a promising support for the physically challenged or elderly society.

IV. Methodology

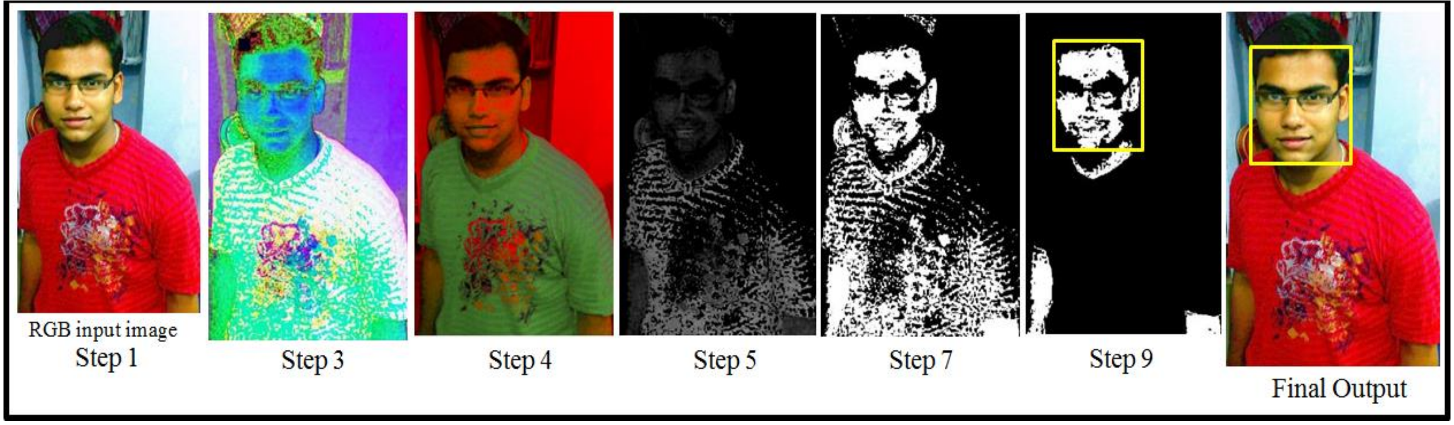


Fig. 3. Pictorial representation of the Steps of Real-time Face Detection and Tracking Algorithm

This is a Software and Hardware based Human System Interaction approach to develop an interactive wireless Robotic interface over Bluetooth connectivity or over Wi-Fi to generate specific industrial applications based on surveillance systems with live biometric authorisation, advanced vehicular facilities, support for the physically challenged and elderly aid.

The User interface involves the Human-Computer Interaction between the User and the Control station which has been accomplished in two different simulation modes – First is with the help of Voice command control, by speech recognition, while the Second is with the help of real-time face detection and tracking by skin tone extraction and analysis. Both the modes have been evaluated in two different wireless embedded platforms individually.

A. System Description

In every Human Computer interaction system Gesture acquisition, analysis and corresponding control signal generation for proper functional execution of the system, forms the fundamental steps of such an implementation. Likewise in our proposed system we have divided the implementation into two different parts. Based on their simulations and evaluated platforms, they are:-

a) Real-time Face gesture controlled Arduino based self-developed Wi-Fi Surveillance Robot.

b) Voice command controlled iRobot Create®.

This section will elucidate the details of both the simulations with their respective platforms.

B. Simulation of Arduino based Self Developed Robot

The design basically focuses on real time face detection and tracking to provide a live biometric authorization to the system. The movements of the detected face is acquired as face gestures corresponding to which different control commands are generated at the control station, which in turn are simultaneously communicated to the embedded prototype through Wi-Fi enabled web access, provided the prototype is within the respective Wi-Fi zone and as a result facilitates the different navigational actions of the embedded prototype using the acquired face gestures. The steps carried out in the design are sequentially illustrated as follows:-

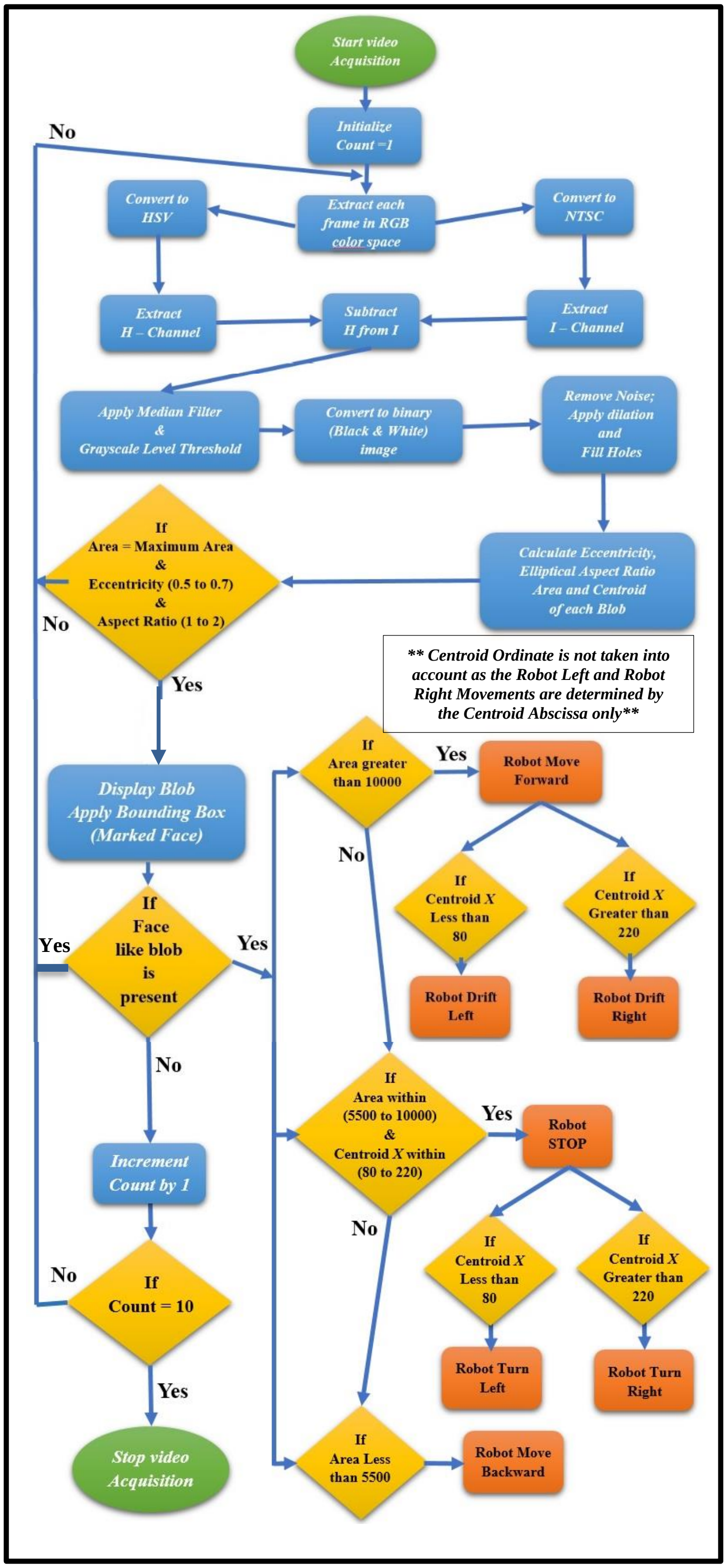
1) Real-time Face Detection

Human face is an identity of an individual which acts as a distinguished and deterministic information for security passwords, database search engines and biometric pattern recognitions. Many approaches have already been made in the field of human face detection for both standalone image frames as well as live video streams since the last decade and presently it has become a major field of interest in current research and technology.

Face detection has achieved moderate detection rates for standalone image frames using complex image processing algorithms which involves color pixel thresholding and normalisation in different color spaces [9], edge based human face detection [10], Skin pixel clustering and quantized skin color regions merging using wavelet packet analysis[11][12], face detection algorithm using eigen image and template matching. These algorithms require huge computational complexity and hence cannot be used for real time processing unless a GPU is involved within the hardware.

In recent trends computer vision based machine learning algorithms proved quite efficient for real-time face detection and tracking among which Viola-Jones algorithm [13] is one of the most popular vision based algorithm for real-time face feature detection as it uses integral imaging and cascade classifiers based on rectangular HAAR-like features to detect the frontal framework of a possible human face like region. However another possible way of detecting and recognising faces would be by mimicking the working principle of human brain with the help of pattern recognition by machine learning algorithms based on training neural networks with supervised learning processes which requires a huge number of training data sets.

Inspired by these works we attempt to come up with a technique which can detect faces with slight tilts and rotations in real-time. Performing experimental trials with several color spaces to remove the maximum number of the non-face pixels, in order to narrow the focus to the remaining predominant skin colored regions [14], we found that the HSV color space fits the best. Since a given picture can have variations of the light incident on it, we had to first make sure that we could cancel the distortions caused by these variations. After going through various color spaces and vision based techniques to reject false positives in real time processing, we decided to go for Skin tone segmentation with a hybrid color space formed by subtracting the Hue channel (H) from the Chrominance channel (I) obtained from the HSV & NTSC color spaces, respectively [15], which results into a skin tone segmented image represented by a 2 Dimensional Matrix. This is then converted to a binary image by grayscale thresholding followed by noise removal and various other morphological processes like dilation and erosion. In order to further classify faces, we implemented a geometric analysis step which would discard any skin tone segmented region that doesn’t fall under the shape of a face and thus reduce the number of false positives. Faces are generally elliptical or rather oval in shape. With this assumption we have filtered the skin segmented regions with an eccentricity filter along with a thresholding in the elliptical aspect ratio i.e. length of major axis by minor axis ratio of each skin segmented blob. The resultant skin segmented blob, with eccentricity ranging from 0.5 to 0.7 and elliptical aspect ratio ranging from 1 to 2, is demarcated as a face like blob which is marked with a rectangular bounding box. This whole process continues iteratively in a loop for successive frames of the webcam acquired live video stream to produce effective real-time face detection.


Start video Acquisition
Initialize Count =1
No
Convert to HSV
Extract each frame in RGB color space
Convert to NTSC
Extract H – Channel
Subtract H from I
Extract I – Channel
Apply Median Filter & Grayscale Level Threshold
Convert to binary (Black & White) image
Remove Noise; Apply dilation and Fill Holes
If Area = Maximum Area & Eccentricity (0.5 to 0.7) & Aspect Ratio (1 to 2)
Calculate Eccentricity, Elliptical Aspect Ratio Area and Centroid of each Blob
No
Yes
Display Blob Apply Bounding Box (Marked Face)
If Area greater than 10000
Yes
Robot Move Forward
No
If Centroid X Less than 80
If Centroid X Greater than 220
Robot Drift Left
Robot Drift Right
Yes
If Face like blob is present
Yes
No
Increment Count by 1
If Area within (5500 to 10000) & Centroid X within (80 to 220)
Yes
Robot STOP
No
If Centroid X Less than 80
If Centroid X Greater than 220
No
If Count = 10
Robot Turn Left
Robot Turn Right
Yes
Stop video Acquisition
If Area Less than 5500
Robot Move Backward

** Centroid Ordinate is not taken into account as the Robot Left and Robot Right Movements are determined by the Centroid Abscissa only**

Fig. 4. Complete Flowchart of the developed Algorithm

2) Real-time Face tracking

This is a computer vision based technology that determines locating faces and their sizes in an arbitrary image and tracks for the face coordinate vectors of successive image frames in a live video stream to achieve a real time live face detection and tracking system. Similarly in our approach the algorithm takes in an RGB image and after performing various steps of skin tone segmentation [16] as discussed above followed by several morphological and filtering processes, it outputs the same RGB image with the largest possible face-like Blob surrounded by a rectangular bounding box[17] as the detected face. When these processes continue iteratively in a loop, the system tracks the user's face by locating the coordinates of the face-like blob within the preview. The rectangular bounding box surrounding the detected face-like blob demarcates the area, instantaneous position vectors and centroid of the blob. The rectangular bounding box moves accordingly along with the movement of the faces in the live preview. The algorithm is developed such that it tracks for the face like blob having the largest area in the preview while discarding other face like blobs in the background. This feature tracks only the user's face regardless of other faces present at the back and hence tracks for the instantaneous face movements of the user within the preview for corresponding Navigation and Control of the Prototype.

3) The complete face tracking algorithm is given as follows:-

Step 1**:** Start the video acquisition in default RGB color frame.

Step 2**:** While frames are being captured initialize Count to unity and do Steps 3 to 10.

Step 3**:** Convert captured RGB frame to HSV color plane and Extract the H channel.

Step 4**:** Convert captured RGB frame to NTSC color plane and Extract the I Channel.

Step 5**:** Subtract HUE (H) channel of HSV from Chroma Intensity (I) channel of NTSC.

Step 6**:** Apply median filter & grayscale level threshold to the resultant image.

Step 7**:** Convert the resultant image to binary black and white image.

Step 8**:** Remove noise, apply dilation and calculate the eccentricity and elliptical aspect ratio of major blob.

Step 9**:** If eccentricity ranges from 0.5 to 0.7 and elliptical aspect ratio ranges from 1 to 2 then mark the Blob as Face.

Step 10**:** Calculate the Area, Location & Centroid of Face-Like Blob within the image.

Step 11**:** If the Parameters of Step 10 are nonzero then repeat Steps 3 to 10, else increment Count by 1.

A cache of 10 iterations is created which checks for the presence of the face like blob within the preview

Step 12**:** If Count is not equal to 10 then do Step 11, else jump out of loop.

Step 13**:** Stop video acquisition.

4) Face Gesture Interpretation and Control

Since the last decade, there have been remarkable advancements in the acquisition of human gestures and their interpretation with the help of computational and mathematical algorithms where any bodily expression or motion is detected by the system and execution takes place accordingly.

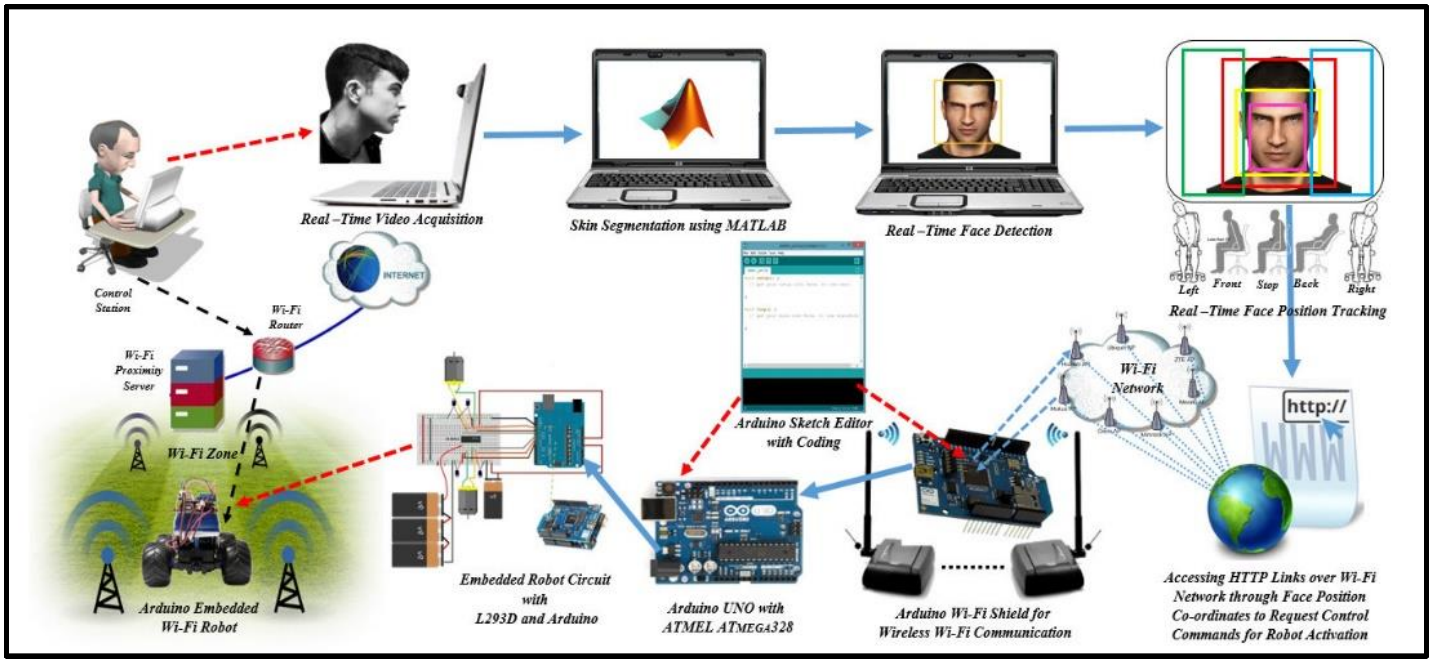

Control Station
Real –Time Video Acquisition
Skin Segmentation using MATLAB
Real –Time Face Detection
Left
Front
Stop
Back
Right
Real –Time Face Position Tracking
INTERNET
Wi-Fi Router
Wi-Fi Proximity Server
Wi-Fi Zone
Arduino Embedded Wi-Fi Robot
Embedded Robot Circuit with L293D and Arduino
Arduino Sketch Editor with Coding
Arduino UNO with ATMEL ATMEGA328
Wi-Fi Network
Arduino Wi-Fi Shield for Wireless Wi-Fi Communication
http://
WWW
Accessing HTTP Links over Wi-Fi Network through Face Position Co-ordinates to Request Control Commands for Robot Activation

Fig. 5. The overall Schematic diagram of the total System Design and Control

Our approach is to develop a hardware and software based integrated system where there is a scope of man-machine interaction remotely without any direct mechanical or hardware interference with the user. The Real-time face tracking algorithm developed is used to detect and locate the user's face in the preview and according to the instantaneous face movements, the algorithm navigates our self-developed embedded wireless prototype in desired directions. The instantaneous face positions are located by recording the coordinates of the rectangular bounding box surrounding the detected face. These coordinates provide a centroid of the rectangular bounding box. The difference in area and centroid of the segmented major face-like blob is used to track the instantaneous facial movements or face gestures for navigation and corresponding control signal generation.

The Navigational Controls to be covered by the Prototype are movements in Forward and Backward directions along with turning Left and Right, both while in motion as well as when at rest. The halt feature, to bring the prototype from motion to rest, is a fundamental control that must be introduced to stop the movements of the prototype. The gestures corresponding to these navigational controls are tracking face movements along the 3-Dimensional Space. The Back and Forth motion of the Face along the Z-axis corresponds to the backward and forward movements of the embedded wireless prototype respectively, considering the PC screen to be the X-Y plane. This back and forth motion of the 2-Dimensional face-like blob is tracked by the difference in area calculations. As the face moves back the area decreases while it increases when the face moves front towards the Webcam. Hence the difference in area (square pixel units) is recorded to simulate the backward and forward motion of the prototype. Tilting left along the negative X-axis and tilting right along the positive X-axis executes the Left turn and Right turn actions of the prototype respectively, while at rest. Similarly we have simulated the turning effects of left and right turns during forward motion by two conditions. First the user must make a forward motion gesture for area increment and then simultaneously tilt left or right of the Y-axis to evaluate the turning effect of left and right turns respectively, while in forward motion. Here the difference in centroid co-ordinates are counted to track the left turn and right turn actions of the Prototype, while the Robot Halt action is marked by the rest position of the User without any movements, when the centroid coordinates of the face like blob merges the line X=0. In this way the navigational controls of the prototype are interpreted at the control station with the help of various face movements.

5) Arduino Based Hardware Platform

The hardware used for prototyping is an open source electronics platform based on a microcontroller board *Arduino UNO R3*, which is designed around a 8-bit Atmel AVR microcontroller ATmega328p (a 28 pin microcontroller) with six analog input pins and 14 digital I/O pins (of which 6 pins can support PWM outputs) along with 6 power pins in addition to it, SDA, SCL, IOREF and AREF pins are also present. The Board features a USB interface together with a DC power jack, an ICSP header and a reset button. Operating at 16 MHz clock speed and 5V input voltage, it consists of 32KB flash memory, 2KB SRAM, 1KB EEPROM and can accommodate various extension boards. It also provides a programming interface and an Arduino Integrated Development Environment (IDE) where programs are written in embedded C or C++ and facilitates a software library which is capable of compiling and uploading programs to the board with a single click. This board uses UART TTL (5V) serial communication to interact with other external peripherals either in the form of USB through Virtual COM Port or through I^2C using the Wire Library supported by the Arduino IDE.

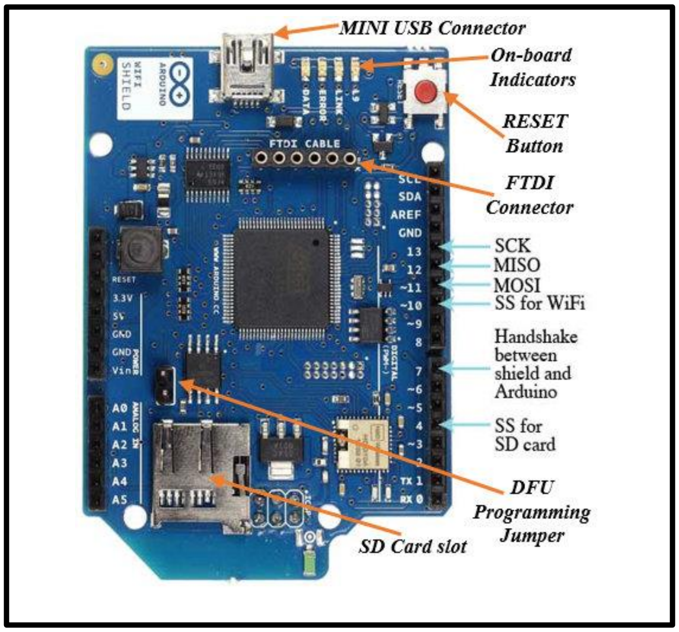


Fig. 6. Hardware Configuration of Arduino Wi-Fi shield

The Arduino Uno R3 board doesn't have any provision to support wireless communication itself or directly get connected to the internet, unless any wireless communication supporting peripheral or any internet accessing device is externally attached and programmed to it.

So in order to connect to the internet wirelessly over Wi-Fi connectivity we have used the *Arduino Wi-Fi Shield* which is compatible to most of the Arduino Boards. Based on the HDG204 Wireless LAN 802.11b/g/n System in-Package, it features a FTDI connector for serial debugging, an SD card slot for downloading/uploading and storing data, a mini USB connector for updating firmware and an AT32UC3 on board chip which provides a network (IP) stack capable of both TCP and UDP data communications through Wi-Fi connectivity. Physically it consists of 30 female pins and six ICSP header pins with long wire wrap headers extending through each of the pins to fit exactly on top of the Arduino Uno board. Communication between Arduino Uno and Arduino Wi-Fi Shield is over SPI bus using Arduino's digital pins 10-13 (CS, MOSI, MISO, and SCLK). While Pin 7 is used for handshaking between the board and the Shield, pin 4 is reserved for SD card storage and hence these pins (4, 7, 10, 11, 12, 13) of the shield cannot be used as Digital I/O pins. The shield can be programmed using the Wi-Fi Library *Wifi.h()* supported by the Arduino IDE and requires the broadcasted SSID name to be connected to Open (unencrypted) networks, together with the Network Password to be connected to WPA personal encrypted networks, while it needs the key index along with the SSID to connect to WEP encrypted networks.

6) *Communication between Control Station and Prototype*

The Control station communicates with the Prototype wirelessly by data communication accessing the internet over a WPA2 (Wi-Fi Protected Access 2) personal encrypted Wi-Fi network. To access internet connectivity both the prototype and the control station must be connected to the same network using its broadcasted SSID and secret Password. The prototype containing the Wi-Fi Shield can serve as either a server accepting incoming connections or a client making outgoing ones. In our approach we have programmed the shield in such a way so that it acts as a virtual Server accepting request from hosts connected to the same network through the default HTTP server Port 80.

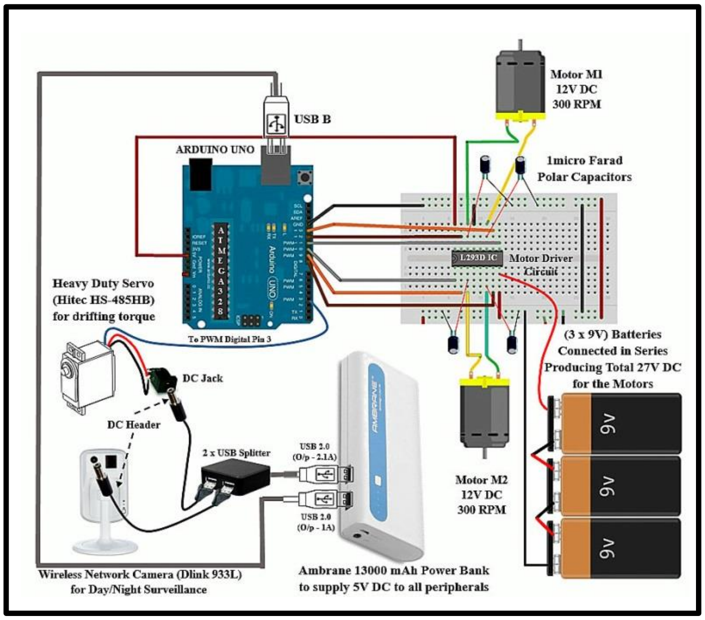


Fig. 7. Total Embedded Circuitry of the Arduino Based Self-developed Robot

In order to connect to a particular encrypted network, the shield first searches for the network in range, when found within range, it sends request to the network for establishing a connection, if all the required network parameters written in the program matches then the network acknowledges the Shield and establishes a connection by providing a local IP address to the shield within the network gateway with which exchange of data takes place between hosts and the shield through the default port 80. An HTML web page is developed at the Shield's URL i.e. the Local IP address assigned to the Shield by the Network. Once gestures are acquired at the control station for a particular navigational control, the control station sends packet data through TCP or UDP in the form of files containing HTTP headers that hits the corresponding link at the HTML webpage with a tagged word written in the code. The HTML web page consists of Hyper Text referenced web links each with a tagged word for all prototype navigational controls that sends control signal files having HTTP headers which request the Wi-Fi shield and the Arduino Uno to activate the prototype via a dedicated motor driving circuit.

7) *Prototype Circuit design and Working Principle*

The hardware design implementation of the prototype includes a motor driver circuit, connected to the microcontroller unit (MCU), that controls the pair of 12V geared DC motors attached to the rear wheels where the Microcontroller Unit sends control signals for forward and backward motions respectively, while the turning effects in left and right directions for both static turnings as well as forward motion drifting effects are generated by a heavy duty servo motor attached to the front wheel.

Our self-developed Prototype is a three wheeled Wi-Fi Robot built on an open source Arduino based embedded platform. Since the MCU outputs 5V DC at an average of 40-60mA as the digital output, it becomes insufficient to drive a 12V geared DC motor at such a power rating. Thus in order to overcome this we need to design a Motor Driver circuit using a 16 pin DIP (Dual in line package) H-bridge IC *L293D*.

The L293D is a motor driver IC which can control two low power DC motors simultaneously, to drive on either direction by just using 4 digital I/O pins. This is a symmetrical IC operating at an average of 5V DC with 4 output pins (3, 7, 11, 14) two on either side, directly attached to each DC motors with an amplified output current rating of 600mA each along with 4 input pins (2, 7, 10, 15) two on either side which are directly connected to the MCU to control the clockwise and counter-clockwise motions of the DC motors by reversing the current directions. The Chip also has a provision of providing an external supply voltage up to 36V DC from which the motors can draw the required amount of voltage to achieve their full RPM ratings. There are two enable pins (1 and 9) on either half of the chip which are responsible for regulating the voltage levels on respective halves and hence can be accessed for speed control of the DC motors attached on either sides.

The Speed Control Mechanism is very essential for special cases where the Motor needs to rotate at minimized or rather variable speeds with respect to its maximum RPM rating at necessary instances. Therefore we have included this feature of speed control mechanism in order to get the effect of geared accelerations of forward motion in our prototype. The implementation of the speed control mechanism is performed by calibrating the maximum RPM ratings of a DC motor into 256 discrete levels from 0 -255. This can be achieved by connecting the enable pins of each half of L293D to two PWM digital output pins of the MCU. Each PWM pin of the MCU by default divides its output voltage range into 8-bits i.e. 256 discrete levels and this calibration can be accessed only with the help of the function *analogWrite()* provided by the Arduino IDE. In our design we have experimentally selected 3 discrete levels out of 256 levels in order to visualize the acceleration effect of the prototype. Hence the speed control mechanism of accelerated forward motion is achieved by calling three user defined functions written in the code of the MCU namely:-

- ✓ *MoveForward()* – the prototype moves at a minimized speed calibrated at the level 150.

- ✓ *MoveForwardFaster()* – the prototype moves at a moderate speed calibrated at the level 200.
- ✓ *MoveForwardSuperFaster()* – the prototype moves at the full RPM rating and highest speed calibrated at highest level 255.

The Gestures corresponding to these speed control levels are similar to the forward motion gestures with a specific increment in area consideration for each level. The prototype moves forward if the blob area is 30% more than default, moves forward faster for 60% more than default , while it moves at its highest rating if the area is 90% more than default. The default area is the area of the face blob of the user at rest position of the robot.

However the physical design of our prototype mainly consists of two rear wheels, each driven by a 12V geared DC motor of 300 RPM rating, with relatively larger radius as compared to the front wheel which is attached to a heavy duty 180 degree servo motor. The rear wheels drives the prototype forward and backward with the servo motor being fixed at 90 degrees which is the default position of the front wheel. The rear wheels are also responsible for clockwise (Right) and anti-clockwise (left) motions i.e. the static right and left movements of the prototype with the servo motor being fixed at 45 degrees and 135 degrees respectively. Connected to a digital PWM pin of the MCU, the main objective to include a servo motor within the prototype is to generate the forward motion turning effects and facilitate the prototype to move along a zigzag path by the user tilting left and right alternatively while performing the forward motion gesture at the control station end. These gestures are tracked by the developed algorithm and the corresponding control signals containing HTTP headers are generated and sent to the Wi-Fi shield in the form of data packets transferred through TCP or UDP over the web via encrypted WPA2 Wi-Fi protocol. The Wi-Fi shield then instructs the Arduino Uno to activate the digital output pins corresponding to the control signal as written in the code and set them HIGH. The digital signals are passed over to the corresponding input pins of the Motor diver IC L293D and the corresponding output pins attached to the DC motors get activated respectively to perform the requested navigational function. The DC motors are externally supplied with a series of 3×9V DC batteries which provide a maximum of 27V DC which is sufficient to drive a pair of 12V geared DC motors at their highest RPM ratings. A 5V DC power bank of 13000mAh is added to the physical setup of the Prototype in order to provide uninterrupted power supply to the MCU, the servo motor and a Surveillance network camera which has been added to the periphery of the prototype in the end. The Surveillance camera is a Wi-Fi based wireless network IP camera, having a facility of infrared Night vision, which accesses the Wi-Fi network of the control station and streams a real-time live feedback of the prototype at the camera URL i.e. the local IP assigned to the network camera by the Wi-Fi network. The user can view the live feed of the camera by logging in to the camera URL with default username and password provided by the manufacturer.

C. Evaluation of Voice Command Controlled iRobot Create®

Speech is one of the oldest and most efficient form of communication that has been used by humans to communicate their feelings. It has been constantly evolving over the time and has risen to become the primary method of communication for us. Hence, a robust speech recognition or automatic speech recognition system is an essential part of any Human-Computer Interactive systems. An embedded speech recognition system, ideally, should be able to isolate background noise and be able to segment and recognize any words spoken to it, be it a single word or a complete sentence. Once recognized, subroutines, based on the recognized words are to be executed that can be used to accomplish various tasks. As can be seen, the system needs to be asynchronous and independent of the other systems. Since the actions are conditional based on only specific voice commands, a speech recognition system, is often without a User Interface.

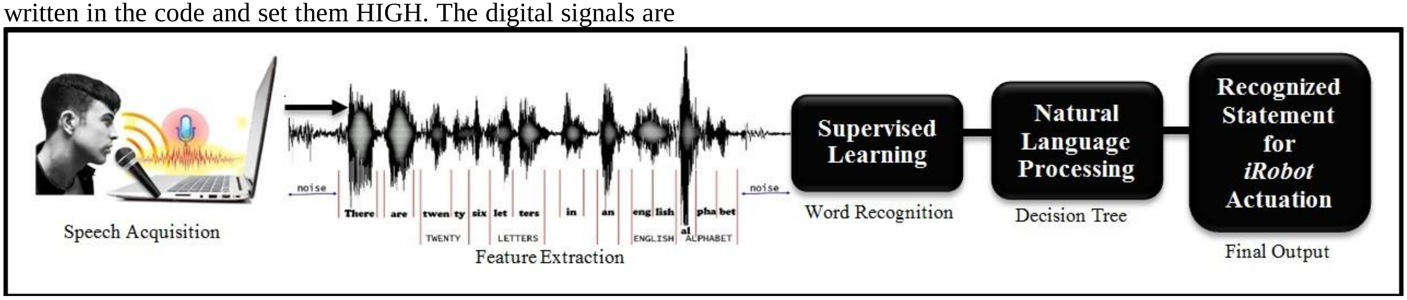

Speech Acquisition
noise
There are twen ty six let ters in an eng lish al pha bet
TWENTY LETTERS ENGLISH ALPHABET
noise
Feature Extraction
Supervised Learning
Word Recognition
Natural Language Processing
Decision Tree
Recognized Statement for iRobot Actuation
Final Output

Fig. 8. Representative Flow Diagram for the Proposed Speech Recognition Algorithm

In case of humans, speech is basically a combination of specific sounds that, in a specific combination, makes up different words, which we have come to associate with different actions and meanings. Moreover, specific combinations of these words in itself, makes up sentences and grammatical structures that fine tune the communication procedure. Humans produce various sounds by varying the shape of the vocal tract, with varying frequencies. Herein comes the problem of the various users and gender differences. Due to different characteristics of properties like pitch, frequencies, etc. sound quality varies widely over genders and even among various persons. This property is used to implement speech recognition as a biometric security measure. Since speech characteristics vary from person to person, the speech can be used for user identification. This property, in turn, gives rise to two types of speech recognition, a user dependent speech recognition and user independent speech recognition. The former, as stated above, is trained to and recognizes speech from only a single user, which is the same as that used in biometric security applications. The latter, however, is a wider, and more complicated implementation of speech recognition, wherein the system recognizes speech irrespective of who the user is. This kind of speech recognition is finding a wider variety of application as it has a tremendous applicability in the field of assistive systems, which is one of the primary fields of Human-Computer Interaction.

In any speech recognition system, there are basically three steps involved in creating a successful system that can give a proper threshold of recognition. The steps involved are:

i. Speech Acquisition.
ii. Speech Analysis, and
iii. Subroutine linking or User feedback.

For the first part, a simple microphone is used to acquire the speech signal. Care should be taken that microphone is unobstructed. Once the signal is acquired, the signal is then analysed. One of the most primitive filtering is the filtering of the background noises and breaking it up into blocks. The blocks are then passed through various pre-processing steps and made ready for matching with pre-trained sets, which are similarly analysed and readied in large datasets. Some of the popular algorithms used in speech recognition are Dynamic Time Warping[18], Hidden Markov Models[19], Neural Networks or Deep Neural Networks[20] and Deep Learning algorithms. Of these, algorithms based on Hidden Markov Models have been proven to be historically robust and to give a very good threshold of speech recognition. Recently however, Deep Learning has been shown to offer good applicability in the field too and some extensive works are being done in the said field.

For our system, we chose to go with the Speech Application Programming Interface or SAPI, developed by Microsoft for the purpose of Speech Recognition and Speech Synthesis. It is a very robust system with a high rate of correct detection and highly accessible programming interface that allowed us to easily code it to work with our system. Since the system is aimed at assistive systems, the ability of speech synthesis is an added advantage. Moreover, being freely distributable, it is available on all windows platforms supporting Windows 98 onwards. The version that we used was SAPI 5 with Visual Basic 2008. The SAPI allows one to recognise predetermined words that are linked to a user defined dictionary. These dictionaries can be dynamically loaded during runtime. But since we aim to make the system a complete and seamless assistive system, we implemented the speech recognition in such a way that the user will speak a natural sentence and the system will be able to recognise what the user wants to command. The intuition is similar to how the human's hearing and comprehension works. Often it has been seen that in order to understand the communication intended behind a spoken sentence, it is not always necessary for the receiver to understand each and every word, but rather an acceptable number of words. For the words that were lost in the communication medium, they can be filled in by pre-emptive filling based on words preceding it. Hence using this intuition, we model a search tree for the Natural Language Processing in our speech recognition system. The SAPI gives us a robust word recognition system which we use to create the probable grammar dictionary. When the user speaks a sentence, our program, using the isolated words, traverses through a pre-modelled search tree. The path of traversal of the tree is then used to determine what the actual intended command by the user is. Once that is determined, a contracted command string is generated which is used by the program to execute the associated sub-actions. For testing purposes we used Visual Basic to send commands to an iRobot Create which is programmed in MATLAB through iRobot Roomba's Open Interface (ROI) protocol. The result was the user just needed to say simple sentences like, "Trace 5 squares" or "Move 10 meters" or "Follow an Object" etc. so as to make the iRobot traverse a square 5 times or to travel 10 meters forward or exactly traverse a path in real time corresponding to the pattern traced by a colored object at the control station end using color based segmentation. However the shapes to be traced has been tested on Squares, Rectangle, Triangle, Circle and the Shape of digit 8 respectively. The User Interface was done in Visual Basic, which can listen for speech in the background asynchronously, thus leaving the system free to execute other functionalities when valid speech is not spoken. The subsequent interfacing of iRobot Create with the control station was done by constant power cycling of the Hardware with MATLAB through a high range Bluetooth Access Module (BAM).

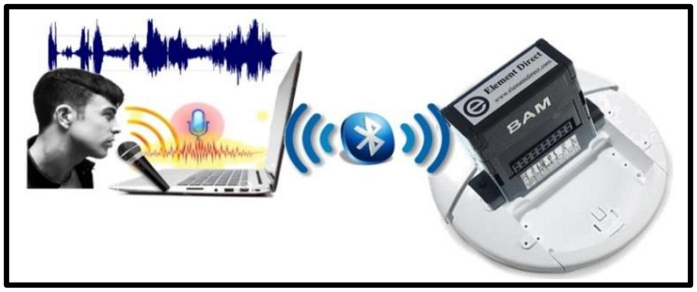
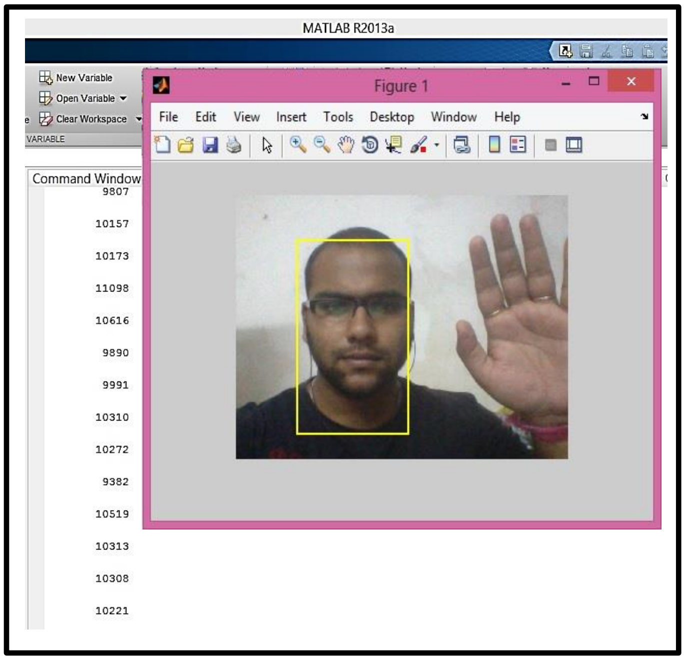


Fig. 9. Setup of Voice controlled iRobot Create over Bluetooth Connectivity



```
9807
10157
10173
11098
10616
9890
9991
10310
10272
9382
10519
10313
10308
10221
```

Fig. 10. Screenshot of MATLAB Command window and webcam preview showing Area Results of Face Tracking Algorithm which establishes a sharp demarcation between Face like regions and Non-face like regions in Real-time

V. Experimental Evaluations and Results

In this paper we put forward a technology which states the concept of controlling embedded wireless systems using two distinct Human intuitions i.e. voice commands and Face movements. Face being the primary element of human identity is a distinct feature (except for the Identical twins) to distinguish between two persons. So detecting and tracking face in real-time with substantial detection results to control various system functions is of primary importance in our approach. Color is the property by virtue of which objects are generally classified. However the face detection by extraction of skin color using color based segmentation is highly dependent on the optimal lighting conditions. Moreover the quality of the acquired image highly depends on the amount of light incident to the aperture of the webcam. The performance of the face tracking algorithm proved to be significantly effective in moderate and ambient light as well as high luminous intensities. It is evident from Fig.10 that the geometrical analysis distinguishes face pixels from non-face pixels quite efficiently. The palm, being classified among the non-face pixels although larger in area than the tested face gets rejected as it doesn't fulfill the criteria of the eccentricity or the elliptical aspect ratio conditions. If there is no face available, even after the cache Count equals to 10, then an alert message appears on the command window "No Face Present" and the video acquisition stops after a while. The screenshot displays the live preview (320×240) video resolution of the real-time face tracking algorithm with corresponding areas of the detected face like blob in square pixels units in MATLAB's command window. TABLE I is generated on the basis of the screenshot result.

TABLE I. Experimental Result Analysis from Figure 10

Iterations	**Area of Face-Like Blob**	**Mean Area**	**Detection Rate**
1	9807	10218.35	95.97%
2	10157	10218.35	99.39%
3	10173	10218.35	99.55%
4	11098	10218.35	91.39%
5	10616	10218.35	96.10%
6	9890	10218.35	96.78%
7	9991	10218.35	97.77%
8	10310	10218.35	99.10%
9	10272	10218.35	99.47%
10	9382	10218.35	91.81%
11	10519	10218.35	97.05%
12	10313	10218.35	99.07%
13	10308	10218.35	99.12%
14	10221	10218.35	99.97%
Mean Area	*10218.35*	Average Detection	*97.32%*

From the above table it can be concluded that there have been 14 iterations out of which the Minimum area is 9382 sq. pixels at the 10th iteration while the Maximum area is 11098 sq. pixels at the 4th iteration.

$$\boldsymbol{Mean\ Area} = \frac{\sum_{i=1}^{14} \boldsymbol{Area(i)}}{\boldsymbol{14}} \ldots\ldots (1)$$

The mean area of the 14 iterations is 10218.35 sq. pixels. So the Detection rate is calculated for each iteration as a percentage deviation from the Mean area.

$$\boldsymbol{Detection\ Rate(i)} = [1 - \frac{\boldsymbol{Abs\{Area(i) - Mean\ Area\}}}{\boldsymbol{Mean\ Area}}] \times 100.... (2)$$

Here the absolute value of [*Area(i)* – *Mean Area*] determines the amount of variation of the data from its tabulated arithmetic mean. According to the equation, when expanded, the magnitude of this variation is subtracted from the *Mean Area* itself, the resultant of which is calculated as a percentage of deviation from the arithmetic (*Mean Area)* for each iteration and hence determine the working consistency of the algorithm for real-time detection of a static face.

But for low lighting conditions the grayscale level threshold drastically changes leading to the rectangular bounding box alternately switching between a non-face skin-like blob and the detected face-like blob for successive frames, which results into increased number of false positives and a higher magnitude of environmental noise. So environmental noise is a multiplying factor which effects detection to a higher extent. After testing in various conditions, as shown in Fig.11, it is highly recommended that bright intensities of hue values of these respective colors (Red, Orange and Yellow) of the visible spectrum should not be present in the background as a major color. It might interfere the performance of the algorithm with poor detection rates.

However the detected and tracked face returns the position vectors of the rectangular bounding box which helps calculate its centroid. The abscissa of the centroid distinguishes between gestures for Left and Right Motion evaluations while the difference in area gives the information for Forward and Backward motions respectively which is evident from TABLE II given below.

To establish Wi-Fi connectivity in between the control station and Self-developed prototype we used a 3G Wi-Fi Dongle (MTS MBlaze). The Wi-Fi shield scans for all the available networks as shown in Fig.12(b) with their respective Signal Strengths and matches for the SSID name written in the code. If match is found then it checks for the encryption type and authenticates the passkeys. Once the connection is established the Wi-Fi shield is assigned with a local IP and the Arduino Serial Monitor asks to browse the IP address in order to access the HTML page of navigational controls for the Prototype activation as in Fig.12(c)

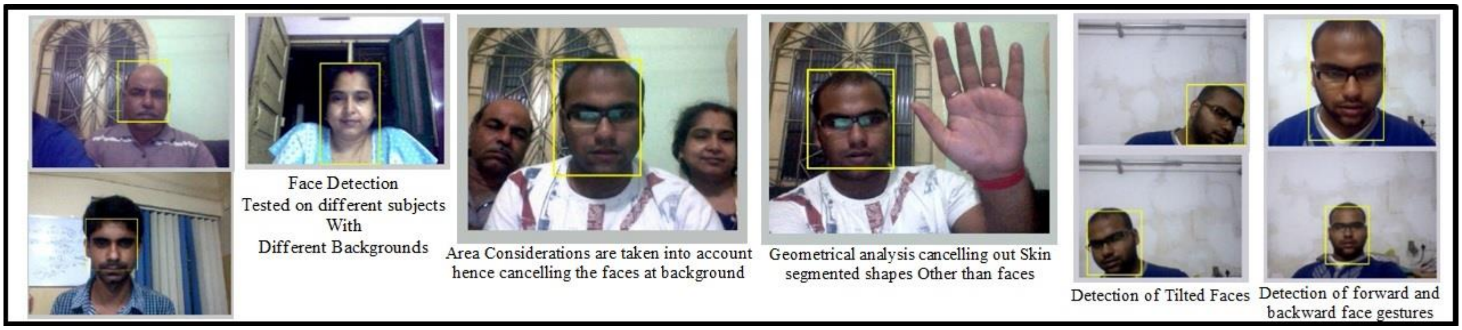
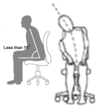
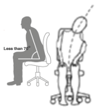

Face Detection
Tested on different subjects
With
Different Backgrounds
Area Considerations are taken into account hence cancelling the faces at background
Geometrical analysis cancelling out Skin segmented shapes Other than faces
Detection of Tilted Faces
Detection of forward and backward face gestures

Fig. 11. Experimental Results of the Face tracking Algorithm

TABLE II. PROTOTYPE ACTUATION WITH CORRESPONDING GESTURE PERFORMANCE

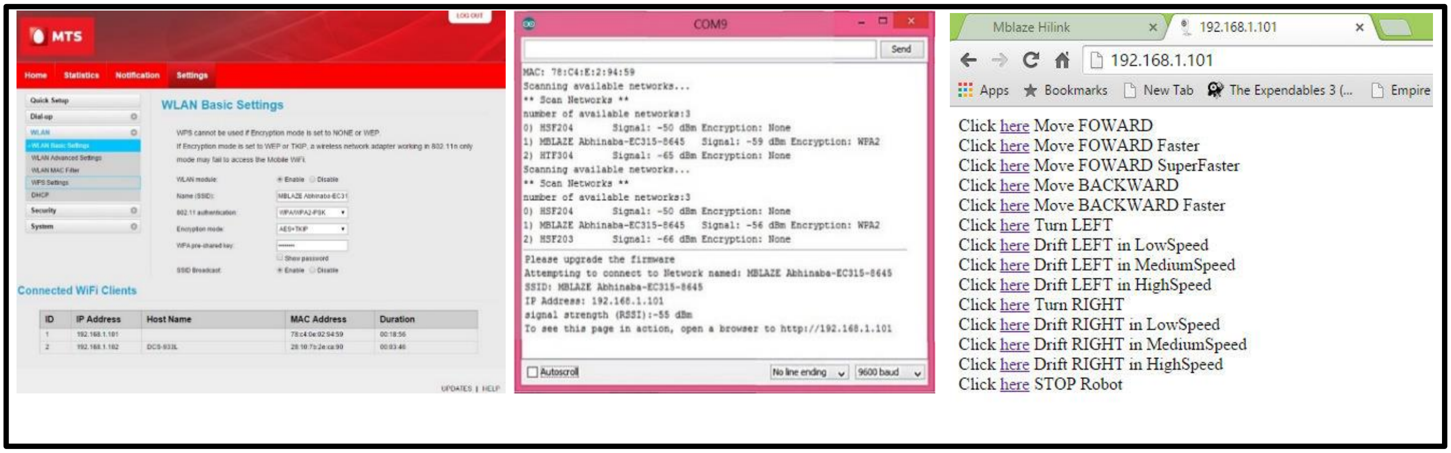
Body Posture	**Face-Gesture Performance**	**Gesture Acquisition**	**Servo Motor Movements**	**DC Motor Movements**	**Prototype Actualization**
		If Centroid abscissa is within (120 to 180) pixels and Area is within (5000 to 11000) Sq.Pixels	Default Position 90 degrees	Both the Motors (M1 and M2) Do not Move	STOP Robot Halt
		If Centroid abscissa is within (120 to 180) pixels and Area is (more than 11000) Sq.Pixels	Default Position 90 degrees	Both the Motors (M1 and M2) Move in the clockwise Direction	↑ Move Forward
		If Centroid abscissa is within (120 to 180) pixels and Area is (less than5000) Sq.Pixels	Default Position 90 degrees	Both the Motors (M1 and M2) Move in the anti-clockwise Direction	↓ Move Backward
		If Centroid abscissa is less than(120 pixels) and Area is within (5000 to 11000) Sq.Pixels	Servo Sweeps to 45 degrees and remains fixed	M1 moves anti-clockwise while M2 moves clockwise respectively	↶ Turn Left from Rest
		If Centroid abscissa is more than(180 pixels) and Area is within (5000 to 11000) Sq.Pixels	Servo Sweeps to 135 degrees and remains fixed	M1 moves clockwise while M2 moves anti-clockwise respectively	↷ Turn Right from Rest
		If Centroid abscissa is less than(120 pixels) and Area is (more than 11000) Sq.Pixels	Servo Sweeps to 45 degrees and returns to default 90 degrees	Both the Motors (M1 and M2) Move in the clockwise Direction	↰ Drift Left
		If Centroid abscissa is more than(180 pixels) and Area is (more than 11000) Sq.Pixels	Servo Sweeps to 135 degrees and returns to default 90 degrees	Both the Motors (M1 and M2) Move in the clockwise Direction	↱ Drift Right

While testing and performing several experimental trials as in TABLE III, it has been found that there is a sufficient amount of lag for a few milliseconds from Gesture performance to command signal generation to the Prototype actuation.

This lag due to network congestion might highly affect real-time applications and may lead to Gesture misinterpretation as well as erroneous results. This lack of synchronization of the system implementation can be overcome by proper systematic network programming, as over here we have just provided a proof of concept for the system implementation.

TABLE III. EXPERIMENTAL OBSERVATIONS FOR PROTOTYPE ACTUATION WITH CORRESPONDING GESTURE PERFORMANCE

Observations	**No. Of Trials**	**No. Of Successful Outcomes**	**Success Rate**
Move Forward	20	19	95%
Move Forward Faster	20	19	95%
Move Forward Super Faster	20	18	90%
Move Backward	20	18	90%
Move Backward Faster	20	17	85%
Drift Left (Low Speed)	20	15	75%
Drift Left (Medium Speed)	20	17	85%
Drift Left (High Speed)	20	16	80%
Drift Right (Low Speed)	20	15	75%
Drift Right (Medium Speed)	20	15	75%
Drift Right (High Speed)	20	17	85%
Turn Left	20	16	80%
Turn Right	20	17	85%
Robot Halt (Stop)	20	16	80%



WLAN Basic Settings

WPS cannot be used if Encryption mode is set to NONE or WEP.
If Encryption mode is set to WEP or TKIP, a wireless network adapter working in 802.11n only mode may fail to access the Mobile WiFi.

WLAN module: ● Enable ○ Disable
Name (SSID): MBLAZE Abhinaba-EC31
802.11 authentication: WPA/WPA2-PSK
Encryption mode: AES+TKIP
WPA pre-shared key: ••••••••
□ Show password
SSID Broadcast: ● Enable ○ Disable

Connected WiFi Clients

ID	IP Address	Host Name	MAC Address	Duration
1	192.168.1.101		78:c4:0e:02:94:59	00:18:56
2	192.168.1.102	DCS-933L	28:10:7b:2e:ca:90	00:03:46

COM9

```
MAC: 78:C4:E:2:94:59
Scanning available networks...
** Scan Networks **
number of available networks:3
0) HSF204       Signal: -50 dBm Encryption: None
1) MBLAZE Abhinaba-EC315-8645   Signal: -59 dBm Encryption: WPA2
2) HTF304       Signal: -65 dBm Encryption: None
Scanning available networks...
** Scan Networks **
number of available networks:3
0) HSF204       Signal: -50 dBm Encryption: None
1) MBLAZE Abhinaba-EC315-8645   Signal: -56 dBm Encryption: WPA2
2) HSF203       Signal: -66 dBm Encryption: None
Please upgrade the firmware
Attempting to connect to Network named: MBLAZE Abhinaba-EC315-8645
SSID: MBLAZE Abhinaba-EC315-8645
IP Address: 192.168.1.101
signal strength (RSSI):-55 dBm
To see this page in action, open a browser to http://192.168.1.101
```

Click here Move FOWARD
Click here Move FOWARD Faster
Click here Move FOWARD SuperFaster
Click here Move BACKWARD
Click here Move BACKWARD Faster
Click here Turn LEFT
Click here Drift LEFT in LowSpeed
Click here Drift LEFT in MediumSpeed
Click here Drift LEFT in HighSpeed
Click here Turn RIGHT
Click here Drift RIGHT in LowSpeed
Click here Drift RIGHT in MediumSpeed
Click here Drift RIGHT in HighSpeed
Click here STOP Robot

Fig. 12(a). Network Home Page Showing SSID broadcast and number of hosts connected

Fig. 12(b). Available networks and the connected Network displayed by the Arduino Serial Monitor

Fig. 12(c). URL of the Wi-Fi device representing the HTML webpage for prototype actuation

However the system implementation aims at several industrial applications of which surveillance security is one major interest. In order to fulfill the purpose we have successfully integrated a network camera into the system which sends the live prototype surveillance feed to the control station wirelessly over the web through a Wi-Fi connectivity. From Fig.12(a) we can see that the Network assigns two different IP addresses to the two Wi-Fi devices connected in which the later one is the network camera with IP address 192.168.1.102 and on browsing this IP address we can access the live surveillance preview of the prototype both in light and dark conditions which is evident from Fig.13.

However to increase the flexibility and security of our implementation we introduced a second type of human intuition to control wireless systems i.e. speech, which solely serves the purpose to navigate and direct systems in a more definite way. TABLE IV. shows the tested results of the speech recognition algorithm. From the above table it is confirmed that the implementation employed here is user dependent i.e. it gives a higher detection rate for the person with whom it has been calibrated. As such a short calibration of the system is essential, albeit a single time calibration. But once the calibration is done, the system achieves a high rate of recognition for the given user even on a wide and varied dictionary of words as well as on complete sentences.

The user dependence, in lieu of our system is also an added bonus. Since, after the initial setup of the system, which includes, among other things, the calibration, the same system can be used to identify the user and serve as a voice print identification. As the implementation is geared towards biometric security based assistive systems or surveillance systems, the system responding to only particular user is an advantage. Hence we decided to go with particular user recognition rather than multi user recognition. Hence based on the recognized statements we have implicitly evaluated certain voice commands which are directly implemented on the iRobot Create for various action execution and can be changed as per necessity.

TABLE IV. EXPERIMENTAL OBSERVATIONS FOR THE SPEECH RECOGNITION ALGORITHM

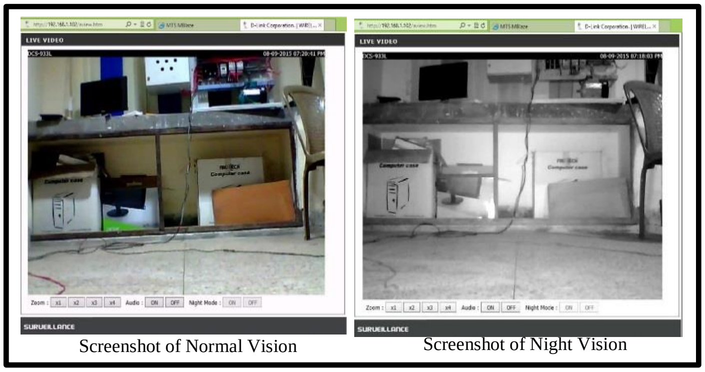
User Dependency	**No. of Trials**	**No. of Successful Recognition**	**No. of False Positives**	**Success Rate**
Particular User	30	26	4	86.67%
Multiple Users	40	13	27	32.5%

VI. DISCUSSIONS AND CONCLUSIONS

From the favorable statistics of results it can be concluded that the implementation of Real-time Face Tracking algorithm is quite efficient with high accuracy rates at moderate and ambient lighting conditions, which overrules many other techniques as it can detect both tilted and rotated faces to some extent with a few limitations. These limitations that affect the detection rates are:-

- ✓ Luminance factor of the Environment with ambient lighting conditions. However, multicolored lights might detect faces but affect the tracking efficiency.
- ✓ Monochromatic Color predominance is another major factor that might affect subsequent detection rates. However, bright intensities of the color spectrum with Hue values ranging from 0 to 45 and 210 to 239 present as a dominant color in the background may produce invariant segmentation where skin tone gets diminished and hence result in false positives.
- ✓ The last but not the least is a hardware limitation where detection depends on the quality of image acquisition that is directly dependent on the aperture width of the Webcam. For real-time processing the more is the aperture width, higher is the quality of the image acquired.

The Arduino based wireless system is considered to be effectively optimized and secured as it provides two modes of control, the first – is a Machine level method by directly accessing the local IP address of the hardware device, while the second one – is through the User-friendly face motion based implementation, both the modes using a protected Wi-Fi connectivity. In contrast to it the user dependent speech recognition based implementation on iRobot Create over Bluetooth connectivity is a more directed and definite form of Human-Computer Interfacing. As for the speech recognition, the user dependency at initial calibration can prove to be a security authorization code for accessing and controlling technologically advanced military armors or weapons. Moreover, the face gestures can be used to control their directions and motions. Similarly, we can integrate wireless systems with these methods for surveillance and monitoring activities, for keeping a check on intruder activities, spying on enemies at Defense Academies or Military Base Camps.



Screenshot of Normal Vision

Screenshot of Night Vision

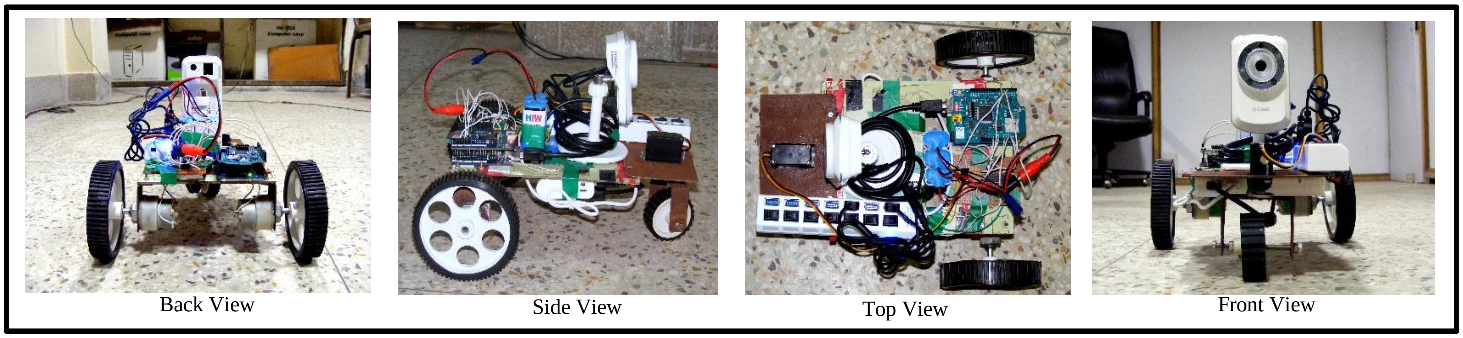
Fig. 13. Screenshot of Network Camera preview in normal and in dark conditions

The live face detection can also act as a live password for accessing surveillance systems in satellite substations. Touchless interfaces are especially useful in healthcare environments where information are accessed while maintaining total sterility. The face tracking algorithm can be used to design assistive monitoring systems at Intensive Care Units in synchronization with other medical instruments where continuous tracking of patient's face, among other things, can provide necessary information about the patient's instantaneous condition. However, if any anomalous face movement is found then, immediate medical assistance would be fetched by the system.

Moreover, the unit configured onto a wheelchair with a camera fitted on the hand-rests focusing the face of the person on it will enable the physically challenged to control motion and navigation of his or her artificial legs comfortably within their domestic limits. Additionally the face gestures performed with different body postures may restrict further physical immobility and might act as a key for the recovery of patients with impaired legs or spinal deformity.

However, these gesture interfaces are also gaining importance in the entertainment fields where touch-free motion based games are being commercialized with various external hardware setups. However, our method can be adopted to implement external hardware free gaming environments with biometric user security where gestures based on face or body movements can be tracked to evaluate various gaming controls with the PC webcam as the only sensor involved.

The proposed work can also be an authoring method capable of operation controlling motions in the fields of industrial automation and control. Vacuum cleaners can be automated by hardware interfacing robots (as in iRobot Roomba) to reduce mechanical effort as well as manpower in various Industrial sites, Resort, Shopping malls, Housing complexes, etc.



Back View

Side View

Top View

Front View

Fig. 14. Arduino based Self-Developed Wi-Fi Robot

As the paper presents an embedded wireless prototype with almost all the features of a geared automobile, so the system design if adopted by the automobile designers can prove to be an effective methodology to integrate real-time gesture controlled automobiles.

The gesture-based access based on face motions (with the user's face being a live biometric authorization) to evaluate various automotive actions in real-time can act as a second mode of control with the geared mechanical access being the primary mode of control.

So as a whole we can conclude that the system implementation is quite a success with the overall experimental results showing high accuracy and detection rates with a wide range of possible advanced commercial implementations as well as industrial applications. Although the Real-time system lag in milliseconds is an interfering factor, it can be eliminated with further research. However, the major advantage of our system over other systems is that it provides real-time face gesture recognition, leading to an effective and natural way for controlling embedded wireless systems.

The proposed speech recognition method is expected to provide effective and implementable solutions for not only just industrial robots but also for intelligent embedded robots like humanoids.

VII. Future Works

The Present System is a real-time gesture tracking system used for motion control of Wi-Fi robots. The face tracking algorithm can be enhanced with a Graphical User Interface (GUI) which presently runs on Standalone scripting codes. The GUI will provide a more user-friendly as well as robust interface for interaction. The face detection algorithm doesn't include a face recognition or Gender detection part, which might make the system much more secured with biometric face recognition for restricted access only. Moreover, the video footage of the Network camera so received can be further processed with various computer vision algorithms to make the robots more intelligent and adaptive to the surrounding environment. Advancements in the speech recognition along with interactive AI as that implemented on CORTANA or SIRI can be used to increase adaptations of the system. Introducing a Real-Time Object Tracking algorithm on the received preview of the network camera would be a breakthrough for such systems and can be extensively used to design assistive robots for elderly aid. Thus, Simultaneous Localization and Mapping (SLAM) can be introduced using the post-processing feature enacted upon the video feed of the network camera. Also, the method can be implemented on advanced automobiles to assist driving and control intellectually.

The Night Vision mode can be used to track objects even at dark provided the Algorithm must be flexible enough and should be tested at various luminous intensities. It can assist elderly drivers with impaired vision or color blindness or night blindness by tracking on road vehicles, traffic signals in real time and thus provide a safe and sound drive by collision detection or avoiding Potholes (especially at Indian roads) even in the dark.

However, these computer vision algorithms provide a more friendly and intellectual interaction of the user with the system. Special robots can be created to assist Room services and housekeeping purposes while accessing them remotely from distant places over Wi-Fi connectivity. Such robots can be widely used in industrial automation and can be implemented for automating cleaning of toilets at public places like shopping malls, railway stations, etc.

The system has a high scope of importance in biomedical applications to assist the physically challenged people. Such systems can be used for maneuvering physical activities of people with paralyzed limbs. If the system is integrated on mobile wheelchairs, the body movements for acquiring face gestures might prove to be a great assistance to partially overcome their physical immobility.

As a whole it can be concluded that the system has a huge scope of further research and application that can prove to be effective in various fields.

Acknowledgement

We would like to thank Arnab Pal, Techno-Marketing Consultant at ESL (e-School Learning) for helping us with the technical details of the project and ESL, www.eschoollearning.net, Kolkata for giving us a platform to work on this project.

References

[1] F. Arce, J. M. G. Valdez, "Accelerometer-Based Hand Gesture Recognition Using Artificial Neural Networks" in *Soft Computing for Intelligent Control and Mobile Robotics Studies in Computational Intelligence*, vol. 318, pp 67-77, 2011.

[2] Luigi Lamberti1 and Francesco Camastra, "Real-Time Hand Gesture Recognition using a Color Glove," *Department of Applied Science, University of Naples Parthenope*, 2010, pp.321-344.

[3] J.S. Kim, C.S. Lee, K.J. Song, B. Min, Z. Bien, "Real-time hand gesture recognition for avatar motion control," *Proceedings of HCI'97*, pp. 96-101,February 1997.

[4] Chao Hy Xiang Wang, Mrinal K. Mandal, Max Meng, and Donglin Li, "Efficient Face and Gesture Recognition Techniques for Robot Control", CCECE, 1757-1762, 2003.

[5] Gesture Controlled Robot using Kinect http://www.e-yantra.org/home/projects-wiki/item/180-gesture-controlled-robot-using-firebirdv-and-kinect

[6] Abhinaba Bhattacharjee, Indrani Jana, Ankana Das, Debasish Kundu, Sudipta Ghosh, Sauvik Das Gupta, "A Novel probabilistic approach of colored object detection and design of a gesture based real-time mouse tracking along with virtual teaching intended for color blind people," *2nd International IEEE Conference on Signal Processing and Integrated Networks (SPIN)*, 19-20 February 2015, Noida, India.,pp.512-519, ISBN: 978-1-4799-5990-7.

[7] Tsang, W.-W.M, "A finger-tracking virtual mouse realized in an embedded system", *IEEE Intelligent Signal Processing and Communication systems (ISPACS)*, 2005,pp. 781-784.

[8] Rafael C. Gonzalez, *Digital Image Processing*, Pearson Education, 3rd ed., 2008.

[9] S Mahanta, S Ghosal, P Das, A Datta, S Debnath, S Das Gupta, "A Novel approach for Graphical User Interface development and real time Object and Face Tracking using Image Processing and Computer Vision Techniques implemented in MATLAB", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 5 (Nov. - Dec. 2013), PP 61-68 www.iosrjournals.org.

[10] Poonam Dhankar and Neha Sahu, "Edge based Human Face Detection using Matlab", Proceedings of IRF International Conference, 16th February 2014, Goa, India. ISBN: 978-93-82702-58-0.

[11] P. Peer, J. Kovac, F. Solina, "Human Skin ColourClustering for Face Detection," EUROCON1993, Ljubljana, Slovenia, pp. 144-148, September 2003.

[12] C. Garcia and G. Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis," IEEE Trans. Multimedia, 1(3), pp. 264-277, 1999.

[13] Viola, Paul and Michael J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. Volume: 1, pp.511–518.

[14] Foley, Van Dam, Foriner, and Hughes, "Computer Graphics: Principles and Practices", *Addison- Wesley, second edition in C*.

[15] Partha Das, Anirban Sarkar, Soumyadeep Halder, Debasish Kundu, Sudipta Ghosh, Sauvik Das Gupta, "A Novel approach towards detecting faces and gender using skin segmentation and template matching," *2nd International IEEE Conference on Signal Processing and Integrated Networks(SPIN)*,19-20 February 2015, Noida, India.,pp.431-436, ISBN: 978-1-4799-5990-7.

[16] Michael Padilla, Zihong Fan – EE368 Digital image processing Project "Automatic Face Detection Using Color Based Segmentation and

Template/Energy Thresholding ", Department of Electrical Engineering, Stanford University, Spring 2002-2003.

[17] Mingli Song, Tao Dacheng, Sun Shengpeng, Chun Chen, S.J. Maybank, " Robust 3D face landmark localization based on Coordinate coding", *IEEE transaction on Image Processing Volume23,* 21 October 2014, ISSN: 1057-7149.

[18] G Poli, J.F. Marie, J.H Saito, " Voice command Recognition with Dynamic Time Warping (DTW) using Graphics Processing Unit (GPA) with Compute Unified Device Architecture (CUDA)," *International Symposium on Computer Architecture and High Performance Computing* ISSN 1550-6533 Pg. 19-25.

[19] A.D Deelip, C.C. Shekhar, "HMM based Intermediate matching Kernel for classification of Sequential patterns of speech using Support Vector Machines", *IEEE transactions on Audio Speech and Language Processing* ISSN 1558-7916 Pg. 2570-2582.

[20] G.E Dahl, Deng Li, Yu Dong, "Context-Dependent Pre-trained Deep Neural Networks for Large-Vocabulary Speech Recognition", *IEEE transactions on Audio Speech and Language Processing* ISSN 1558-7916 Pg. 30 -42.

The Effect of Feature Selection on Phish Website Detection

An Empirical Study on Robust Feature Subset Selection for Effective Classification

Hiba Zuhair[a]
Dept. of Computer Science
Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM, Johor Bahru, Johor, Malaysia;
Al-Nahrain University, Baghdad, Iraq

Ali Selmat
UTM-IRDA Center of Excellence
Universiti Teknologi Malaysia and Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM, Johor Bahru,
Johor, Malaysia

Mazleena Salleh
Dept. of Computer Science
Faculty of Computing, Universiti Teknologi Malaysia,
81310 UTM, Johor Bahru, Johor, Malaysia

Abstract*—Recently, limited anti-phishing campaigns have given phishers more possibilities to bypass through their advanced deceptions. Moreover, failure to devise appropriate classification techniques to effectively identify these deceptions has degraded the detection of phishing websites. Consequently, exploiting as new; few; predictive; and effective features as possible has emerged as a key challenge to keep the detection resilient. Thus, some prior works had been carried out to investigate and apply certain selected methods to develop their own classification techniques. However, no study had generally agreed on which feature selection method that could be employed as the best assistant to enhance the classification performance. Hence, this study empirically examined these methods and their effects on classification performance. Furthermore, it recommends some promoting criteria to assess their outcomes and offers contribution on the problem at hand. Hybrid features, low and high dimensional datasets, different feature selection methods, and classification models were examined in this study. As a result, the findings displayed notably improved detection precision with low latency, as well as noteworthy gains in robustness and prediction susceptibilities. Although selecting an ideal feature subset was a challenging task, the findings retrieved from this study had provided the most advantageous feature subset as possible for robust selection and effective classification in the phishing detection domain.

Keywords—phish website; phishing detection; feature selection; classification model

I. Introduction

Phishers impersonate trustworthy websites of financial organizations through online transactions. Many efforts have been made to overcome the phishing attacks through numerous phishing detecting approaches. Nevertheless, phishing has caused enormous money loss in the cyberspace over the past years, which has motivated researchers to seek effective phishing detection techniques that protect users' digital identity [1-3]. In general, phishing detection techniques fall into several categories due to the deployed scenarios of detection. In the literature, Islam & Abawajy [4] roughly categorized them into non-classification and classification techniques. Specifically, white lists of famous trustworthy URLs; black lists of valid phish URLs; heuristics; and information flow techniques were categorized as non-classification techniques. In contrary, classification techniques involved those relied on machine learning classifiers and data mining based scenarios. They differ in terms of classification accuracies, rates of classification errors, and demands on external resources [1-5]. However, they commonly have deployed features as the key factor for classification task, such as hybrid features. Besides, classification task mostly rely on extracting a set of features from tested instances (i.e. emails and websites) and deploy them to distinguish phish instances from the legitimate ones [1-5]. Thus, classification techniques outperformed their competitors by intuitively detecting phishing that exploits the web to protect clients [3, 6]. Moreover, they could automatically extract features from webpage content; URL of websites, hosting information, and classifying their phishness [7 and 8]. Besides, the usage of hybrid features supported the generality of the classification techniques to classify phishing variations and such techniques reported high rates of detection accuracy than those provided by their competitors [4, 6 and 9]. However, constraints like high-dimensionality of feature set, hybridity of features, their irrelevance to the corresponding classes (i.e. phish and legitimate), their dependency on each other, their redundancy on the examined feature space, and heterogeneity of their values (i.e. discrete and continuous values) might degrade detection accuracy. In addition, they might have increased the false detection errors and computational costs. Then, they would limit the overall effectiveness of classification techniques in the real-world experience along with their scalability to the enormous web data and the evolving phish exploits [5, 9].

Hence, to tolerate with the aforesaid issues, researchers had looked into their constructed classification models via feature selection methods that played an important role in data analysis during the classification task. Such methods typically refined the extracted set of features into a minimal and effective subset for the classification task. Besides, they eliminated the least representative features by applying the lowest discrimination on the tested data. However, these assisted methods yielded different outputs of feature selection. Meanwhile, as for the existing researches; specifically in phishing websites detection, the direct comparison of such differences had been neglected. In their evaluations, they underlined the differences with respect to the detection accuracy and overall performance [4-12]. They rarely quantified feature selection methods in terms of (i) the measure of feature's prediction susceptibility that they had utilized, (ii) their scalability under different feature sets' dimensions, (iii) the goodness of their output in the presence of different classification models, (iv) the stability of their output against evolving data and phishing variations, and (v) the similarity between the outputs of multiple feature selection methods.

Besides, the causality between the aforesaid issues and the optimum choice of feature selection subset had been highlighted. It quantified the highest quality of selected feature subset that yielded the best case of detection accuracy with least error rate as possible. Moreover, this contribution is extended by testing the selected feature subset across multiple classification models. Apart from that, this study promotes its contribution by handling a proposed set of hybrid features. Hence, it is hoped that the proposed features, the characterized literatures, the highlighted issues, and the empirical tests would offer a global picture on phishing detection assisted by feature selection. Moreover, they could be regarded as the baselines for future works to appropriately choose the feature selection methods for their classification models.

In this context, this study characterizes the prior works, and critically appraises them with respect to their frontiers in feature selection as presented in Section II. Then, Section III recommends certain criteria and depicts their relevant terminologies to assess both resilience and effectiveness of selective feature subsets. Section IV, practically appraises feature selection exploits and testifies their outcomes in the presence of the recommended criteria. Based on the stated findings, Section V deduces the present work on hand and gives an outlook to the future implications.

II. Background

A. Feature Selection Methods

All feature selection methods aim at reducing the dimensionality of the feature space and in enhancing the compactness of the features. Meanwhile, in data processing, specifically data mining and machine learning approaches; a large number of features may cause problems of high dimensionality, irrelevance, and redundancy [13]. Therefore, in order to reduce the dimensionality and to obtain the most representative features that could effectively predict instances over a given dataset, data pre-processing is needed [13 and 14]. Mainly, feature selection has been considered as a data pre-processing technique that chooses a minimum subset of *m* features from an original set of *n* features. Accordingly, the selection involves: a search procedure for feature subset generation, and an evaluation criterion for iterative feature selection [13, 14]. Furthermore, the search procedure often discards or adds one feature based on its evaluation outcome, whereas the evaluation criterion compares that feature with the previously selected one regarding to either its information, or dependency, or consistency, or distance or its transformation. However, feature selection methods differ in specifics and parameters that can be tuned for both the search procedure and the evaluation criterion [13, 14]. *Table I* enlists four feature selection methods that had been adopted for phishing detection in the reviewed literature, which were characterized by search procedure, as well as evaluation specifics and criteria.

TABLE I. CHARACTERIZATION OF FEATURE SELECTION METHODS (ADOPTED FROM [13-15])

Feature Selection Method	**Search Procedure**	**Specifics**	**Evaluation Criterion**
Information Gain (IG)	Filter	Information	$IG(S,a) = Entropy - \sum_{V\in a}\frac{\lvert S_V\rvert}{\lvert S\rvert} * Entropy(S_V)$ (1) "Where *S*, S_V, *V* and *a* are the collection of instances, a subset of instances with *V* of a, a relevant value and an attribute, respectively."
Correlation Based Feature Selection (CFS)	Filter	Consistency	$p(C = c\|V_i = v_i) \neq p(C = c)$ (2) "Where, V_i is said to be relevant if there exists some v_i and c for which $p(V_i = v_i) > 0$."
Chi-squared (χ^2)	Filter	Transformation	$\chi^2 = \frac{N\times(AD-CB)^2}{(A+C')\times(B+D)\times(A+B)\times(C+D)}$ (3) Where$A = \#(t,c)$,$B = \#(t,\neg c)$,$N = A + B + C + D, = \#(\neg t, c)$, $D = \#(\neg t, \neg c)$, and t and $r$are independent parameters
Wrapper Feature Selection (WFS)	Embedded with Classifier	Accuracy	Greedy search for feature subset in a forward selection and backward elimination of features

B. Related Works

At present, vast literature is available on the merits and demerits of phishing detection campaign. Towards devising anti-phishing solutions for the specific problem at hand (i.e. phishing websites), many proposals have been introduced and experiments conducted by using different machine learning-based approaches combined without features extraction and features selection. For instance, Likarish et al. [15] developed a Bayesian filter to identify phish websites based on retrieved tokens obtained from the HTML document and constructing DOM (Document Object Model) with the aid of DOM parser. Then, researchers at Google Inc., Whittaker, Ryner & Nazif [16]; worked on the up-gradation of Google's phishing blacklist integrated with a classifier. In addition, another anti-phishing technique was developed by Bergholz et al. [17] to phish email filtering by analyzing several extracted features related to body, external, and model based on examined emails. The developed techniques involved two training phases; one for model-based features and the other was for the rest of the features. Later, CANTINA$^+$ was proposed by Xiang, Hong, Rose, and Cranor [18] with three classifiers and ten features derived from the URLs and the contents of webpages, as well as some online features for highly accurate results of phishing detection. Meanwhile, Zhang Liu, Chow, and Liu [19] introduced a linear classifier *Naïve Bayes* (*NB*) in order to detect eight textual and visual features on suspected websites for phishness prediction. The used classifier returned a normalized number; reflecting the likelihood of the suspect website as being phished or legitimate. Likewise, a *Supervised Machine Learning* (*SVM*) classifier was developed by He et al. [8] to predict phishness on examined webpage by exploiting webpage identity and some textual features. The textual features were extracted by using a well-known information retrieval method to be deployed for classification process. Contrarily, a phish webpage detector was proposed by Li, Xiao, Feng, and Zhao [20] based on visual features and DOM objects of the webpage content that learned and tested over datasets by using Semi-Supervised Machine Learning (*TSVM*) classifier. Furthermore, Kordestani and Shajari [21] applied three classifiers, including Naïve Bayes (*NB*), *Supervised Machine Learning* (*SVM*), and *Random Forest* (*RF*), on a randomly selected dataset to predict phishness in suspected websites. They were deployed for phishness prediction with the presence of URL and online features. Then, Gowtham and Krishnamurthi [22] extracted fifteen, which were trained by using *Supportive Vector Machine* (*SVM*) classifier and a whitelist through two modules. The first module involved checking the identity features of the examined website against a pre-defined white list of legitimate ones, whereas the second module predicted phishness of the examined webpage based on its login form features via *SVM* classifier. However, the application of the aforesaid proposals encountered some trade-offs related to the processing of large and realistic datasets, the extraction of hybrid features, the analysis of their heterogeneity, increasing storage requirements and processing time, as well as some costly miss-classifications.

Moreover, it is worthy to mention that final decisions of phishing detection relied potentially on predictive features against phishing susceptibility. More precisely, phishing detection in the presence of predictive features should yield minute amounts of both valid phish misclassifications and losses of valid legitimate instances. Thus, researchers were motivated to maintain some feature selection methods as those briefly described in *Table II* to cope with the aforesaid factors. In the literature, Pan and Ding [23] proposed phishing detector based on applying *Supportive Vector Machine* (*SVM*) classifier and extracting both textual and Document Object Model (*DOM*) features from the examined webpages. They employed two major components for their detector, including an information retrieval strategy to extract textual features and *Chi-squared* (χ^2) criterion to select the most effective features. Then, Ma Ofoghi, Watters, and Brown [24] experimentally analyzed seven webpages and pages to rank the features with the aid of a filter-based feature selection method, *Information Gain (IG)*, to phish website classification and deploy two classifiers that varied in their classification accuracy due to the selected features. On top of that, Khonji, Jones, and Iraqi [25] enhanced classification performance by selecting the most effective subset of the most commonly used 47 features. Both filter-based and Wrapper-based feature selection methods, such as *Information Gain* (*IG*), *Correlation Based Feature Selection (CFS)*, and *Wrapper Feature Based Selection (WFS)*, were developed with machine learning classifiers to predict phish emails. The classification results differed due to the employed feature selection method and the number of selected features. On the other hand, Basnet, Sung, and Liu [26] analyzed high dimensional feature space, including 177 features extracted from both the content and URL of websites to select the best feature subset. In fact, several subsets were considered for application of *Wrapper Feature Based Selection* (*WFS*) and *Correlation Based Feature Selection* (*CFS*). They were trained over a dataset with the aid of *Logistic Regression* (*RF*) classifiers. Nevertheless, they varied in selecting the most contributing features such that classifiers caused variation on detection accuracies. Later, Zhang, Jiang, and Kim [27] developed automatic detection approach for Chinese e-business websites by incorporating the unique features extracted from URL and contents of website. Alongside, Hamid and Abawajy [28] proposed a multi-tier detector to phish emails filtering with the aid of Adaboost and SMO classifiers in an ensemble design. Moreover, they used *Information Gain (IG)* and clustering strategy to quantify the best predictive features of phish emails and also tested the outcomes over three large scale datasets. However, large size dataset, imbalanced datasets, redundancy, the limit of cluster size, and error rates emerged as the key issues in their work.

C. Shortages

In order to offer a global view on feature selection for exploitation in phishing detection domain, *Table II* characterizes the previous works with respect to their deployed feature selection methods and their limitations.

TABLE II. RELATED WORKS WITH LIMITED FEATURE SELECTION METHODS

Citation	Feature Selection Method (S)	Classifier (S)	Related Limitations
Pan and Ding, 2006 [23]	χ^2	SVM	▪ Heterogeneity of features values ▪ Dissimilarity of selection outputs ▪ Computational cost ▪ Redundancy and irrelevance
Ma et al., 2009 [24]	IG	C4.5	▪ Heterogeneity of features values
Khonji, Jones and Iraqi, 2011[25]	IG, WFS, CFS	RF	▪ Dissimilarity of selection outputs ▪ Imbalanced Data ▪ No scalability
Basnet., 2011 [26]	CFS, WFS	LR, RF, C4.5	▪ Computational cost ▪ Dissimilarity of selection outputs ▪ Heterogeneity of features values ▪ Redundancy and irrelevance
Zhang, Yan and Jiang, 2014 [27]	χ^2	SMO, LR, NB	▪ Redundancy and irrelevance
Hamid and Abwajy, 2014 [28]	IG	Adaboost, SMO	▪ Heterogeneity of features values ▪ Non-scalability ▪ Computational cost ▪ Dissimilarity of selection outputs

As depicted in *Table II*, the surveyed works often deployed sub-optimal feature subsets for phishing detection due to some limitations. Such limitations include: the dependency of feature selection outcomes on a given dataset, different feature selection outcomes across different classification models, heterogeneity of features values, and un-scalable feature selection method to more challenging datasets [23-28]. Furthermore, most of the dedicated efforts focused on discarding the relevant features rather than the redundant ones during feature selection [23-28]. Besides, since they are mutually dependent on other features belonging to the same targeting class; the redundant features might distort the classification task and then degrade its accuracy by producing high error rates [29 and 30]. Consequently, *Table III* underlines some striking issues like non-scalability, heterogeneity, non-robustness, irrelevance, and redundancy that must be considered to deal with feature selection limits [29-31].

TABLE III. STRIKING ISSUES OF FEATURE SELECTION, ADOPTED FROM [29-31]

Striking Issues	Description
Non-scalable Feature Subset [29]	The deployed features rarely raise the classification accuracy to the best case as possible under different selection scenarios and over different datasets.
Redundant Features [29, 30]	Since the high-dimensional data have a substantial amount of irrelevant features which require high computational cost selection strategy to reduce. Such strategy potentially causes inefficient classifier. Irrelevant features, in turn, may contain redundant and non-redundant features which require a robust feature selection strategy capable to handle their redundancy.
Irrelevant Features [29, 30]	Large scaled and realistic datasets like that involved in anti-phishing techniques the may contain high fraction of irrelevant features. Because of the exponential growth of more sophisticated and deceptive phishing features, the resultant irrelevant features highly degrade the classifier's performance.
Feature Values Heterogeneity [31]	Websites are inconsistent datasets with various hybrid features that have different values - discrete, categorical and continuous values. For any collected dataset, the extracted hybrid feature space is heterogeneous in values and huge in size. That is, in the presence of any extracted or selected subset of features, the machine learning classifier should be able to categorize them for both training and testing purposes with a minimum loss of feature values.
Non-robust Feature Subset [31]	When applying feature selection for knowledge discovery, robustness of the feature selection result is a desirable characteristic, especially if subsequent analyses or validations of selected feature subsets are costly. Modification of the dataset can be considered at different levels: perturbation at the instance level (e.g. by removing or adding samples), at the feature level (e.g. by adding noise to features), or a combination of both.

III. ASSESSMENT MEASURES

Other than that, as for the problems at hand (*Table III*), the outcomes of selective feature subset must be quantified on its scalability, goodness, stability, and similarity over multiple datasets [29-33]. In addition, the assessment of outcomes prediction susceptibility against phishing over different datasets is a noteworthy issue to be highlighted towards obtaining the most advantageous features [34]. Thus, specific measures adopted by prior researchers in different fields have been recommended in this work (*Table IV*) to test and to assess the outcomes of feature selection methods [31-37]. Such measure can be considered as comparison baselines for any further study on feature selection effects.

TABLE IV. RECOMMENDED EVALUATION MEASURES FOR FEATURE SELECTION [32-38]

Metrics	Advantage	Evaluation Criterion
Goodness [31]	It measures how well the selected feature subset can accurately classify extremely imbalanced datasets.	$Goodness(S_i) = \frac{1}{Y}\sum_{i=1}^{Y}\frac{N_i^{tp}}{N_i}$ (4) Where Y, N_i^{tp} and N_i are the number of classes in the dataset, the number of true positive of each class and the total number of instances for class *i* respectively
Stability [31, 32]	It quantifiably proves whether the selected features are relatively stable against variations of real world datasets over a period of time.	$Stab(S) = \sum_{f_i \in X}\frac{F_{f_i}}{N} \times \frac{F_{f_i}-1}{\lvert D\rvert-1}$ (5) Where $f_i \in X$ and $\frac{F_{f_i}}{N}$ are all features in a collection dataset S and the relative frequency of each feature in a subset. If all subsets are identical then *Stab(S)* is close to 1; otherwise is close to 0.
Similarity [31, 33]	It compares the behaviour of multiple feature selection methods and their selected features on the same data.	$Sim(t_1, t_2) = 1 - \frac{1}{2}\sum\left\lvert\frac{F_{f_i}^{t_1}}{N^{t_1}} - \frac{F_{f_i}^{t_2}}{N^{t_2}}\right\rvert$ (6) Where $F_{f_i}^{t_1}$ and $F_{f_i}^{t_2}$ denoting the number of frequencies of feature f_i in two candidate feature selection methods t_1 and t_2 respectively. Similarity takes values within[0,1].
Prediction Susceptibility Or Phishness Ratio [34]	A ***phishness ratio*** restates the prediction susceptibility of selective feature set to phishing upon each instance in the dataset. The probability $P_r(P \mid t_i)$ of estimated phishness along with a feature $t_{i.}$ is computed across all instances in the dataset. Then, the instance's phishness is computed by averaging the probability of all its related features.	$P_r(P \mid t_i) = \frac{N_{t_i \to P}}{N_{t_i \to P} + N_{t_i \to L}}$ (7) $Phishness(S) = \frac{\sum_{i=1}^{n}\Pr(P \mid t_i)}{n}$ (8) Where S is the examined webpage, *Phishness (S)* is the prediction of phishing susceptibility, t_i is the feature in *S*, $N_{t_i \to P}$ is the number of occurrences of t_i in phish instance, $N_{t_i \to L}$ is the number of occurrences for t_i in legitimate instance. and *n* is the number of features in *S*.
Minimal Redundancy [35, 36]	It eliminates duplicate features that having another one replicate them in the dataset.	$Min\, R(S) = \frac{1}{\lvert S\rvert^2}\sum_{x_i, x_j \in S} I(x_i, x_j)$(9) Where *R(S)* is the set of highest mutually exclusive features that selected between x_i and x_j.
Maximal Relevance [35, 36]	It selects most relevant features to the target class and highly affecting the classification output.	$\text{Max}\, D(S, c) = \frac{1}{\lvert S\rvert}\sum_{x_i \in S} I(x_i, c)$(10) Where *D(S,c)* is the mean value of all mutually informative features x_i with respect to class *c*.

mRMR* $\boldsymbol{\Phi(D, R)}$ *[37]	This criterion selects a subset feature compactness composed of the most relevant and least redundant features from the original set simultaneously.	$Max\,\Phi(D, R), \Phi = D - R.$ (11) Where, *D and R* indicate the dependency between a feature x_i and its class, and the highest relevance between features x_i and x_j in the same feature set.

IV. Empirical Test and Discussion

Based on the recommended measures presented in *Section III.C*, the empirical test was conducted to state not only the variations of assisted feature selection methods on *Prediction Susceptibility*, *Goodness*, *Stability*, *Similarity*, and *Scalability*, but also it assessed outputs of the simultaneous discarding criterion of redundant and irrelevant features (*mRMR*). To the best of our knowledge, this type of empirical test with the aid of the recommended criteria is scarcely underscored in the literature of phishing detection despite of its significance for feature selection. Hence, an empirical test was implemented on a specific test-bed that was set to extract a large number of hybrid features. Then, a comparison was made on the effectiveness of the best chosen feature subset across different classification models. Test-bed is described, results are reported, and discussion is summarized in the following:

A. Test-Bed and Features

A wide range of aggregated phish and legitimate webpages were considered as test-bed for this study. Mostly they are reported in public archives such as *PhishTank*, *CastleCops*, and *Alexa*. Both *PhishTank* and *CastleCops* are phishing data archives that volunteers frequently update them with valid living phish webpages. While, *Alexa* archive is publicly used to retrieve valid legitimate webpages. We chose such archives because they were commonly used by prior researchers in the literature of phishing detection [15-28]. *Fig. 1* illustrates the aforesaid test-bed in terms of dimension, the number of phish webpages and the number of legitimate webpages. In *Fig. 1*, the test-bed consists of three multiple datasets: Dataset1, Dataset2 and Dataset3. Dataset1 composed of 1000 webpages, Dataset2 composed of 5000 webpages but Dataset3 consists of 10000 webpages. Multi-dimensional test-bed helped to empirically assess the outcomes of the reviewed feature selection methods towards demonstrating the most suitable one among them for phish website detection. Indeed, the webpage content and URL can be used to characterize each instance included in the aforesaid datasets such that they can be categorized accordingly to a specific class either phish or legitimate.

Consequently, the characterized datasets with their features and corresponding classes helped to generate the required feature space. *Fig. 2(a)* illustrates the structure of the generated feature space in terms of class label, feature index, the feature itself, and its value. Furthermore, *Fig. 2(b)* shows a part of the database schema to provide a global view on how raw data could be generated. Moreover, a set of web development tools, such as Fireburg, Jsoup, and Import.Io, had been helpful in implementing this task. Besides, several publicly used tools, such as KNIME and WEKA -the Waikato Environment for Knowledge Analysis, were employed for feature selection implementation and tests.

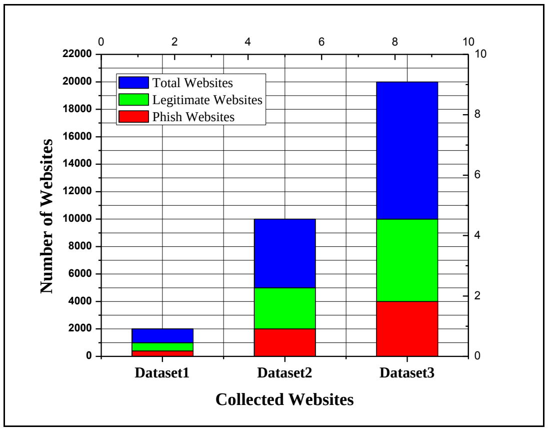
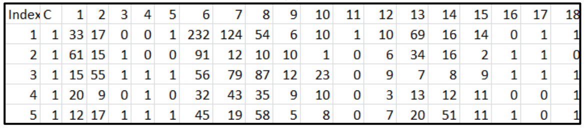


Fig. 1. Description of collected datasets in terms of the total number of instances, legitimate websites and phish websites

	Feature Vectors	*Class Label*	*Features*				
	$\boldsymbol{i}$	$\boldsymbol{C_j}$	$\boldsymbol{f_{j,1}}$	$\boldsymbol{f_{j,2}}$	$\boldsymbol{f_{j,i}}$	…	$\boldsymbol{f_{j,n}}$
	$\boldsymbol{j}$		$\boldsymbol{v_{j,1}}$	$\boldsymbol{v_{j,2}}$	$\boldsymbol{v_{j,i}}$	…	$\boldsymbol{v_{j,n}}$
m Samples	$\boldsymbol{W_1}$	C_1	$v_{1,1}$	$v_{1,2}$	$v_{1,i}$		$v_{1,n}$
	$\boldsymbol{W_2}$	C_2	$v_{2,1}$	$v_{2,2}$	$v_{2,i}$		$v_{2,n}$
	$\boldsymbol{W_j}$	C_j	$v_{j,1}$	$v_{j,2}$	$v_{j,i}$		$v_{j,n}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	$\boldsymbol{W_m}$	C_m	$v_{m,1}$	$v_{m,2}$	$v_{m,i}$		$v_{m,n}$

(a)



Index	C	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	33	17	0	0	1	232	124	54	6	10	1	10	69	16	14	0	1	1
2	1	61	15	1	0	0	91	12	10	10	1	0	6	34	16	2	1	1	0
3	1	15	55	1	1	1	56	79	87	12	23	0	9	7	8	9	1	1	1
4	1	20	9	0	1	0	32	43	35	9	10	0	3	13	12	11	0	0	1
5	1	12	17	1	1	1	45	19	58	5	8	0	7	20	51	11	1	0	1

(b)

Fig. 2. Illustrative example of (a) generated dataset structure and (b) database schema

In *Fig. 2(a)*, the j^{th} webpage is characterized as a vector of features W_j . Then, all feature vectors extracted from m-dimensional set of webpages are represented as combined together in a feature matrix M such that $M = \{W_1 \quad W_2 \quad W_m\}$; where m indicates the number of feature vectors included in M. Each entry vector W_j in M consists of its features' indexes and their corresponding values along its corresponding class label as the first column, i.e. $W_j = \{C_j, (f_{j,1}, v_{j,1}), \quad (f_{j,2}, v_{j,2}), \dots \quad (f_{j,n}, v_{j,n})\}$; where n is the number of features, $f_{j,i}$ is the index of each i^{th} feature of j^{th} feature vector W_j, where $0 \le f_{j,i} \le 1$, $i = 1, 2, 3, \dots, n$ and$j = 1,2 ,3 , \dots, m$. Whereas C_j is the label of the class such that $C_j \in \{1, 0\}$ with $C_j = 1$ and$C_j = 0$, which indicates the membership of W_j in the phish class or in the legitimate class based on its corresponding features [38, 39] as portrayed in *Fig. 2(b)*. Further, features of Boolean values are mapped into either 0 or 1, and features of Continuous quantities are represented as numeric quantities. *Appendix I* enlists the original set of features extracted from all webpages included in the test-bed. Totally 58 features were included in the original feature set. 48 features were extracted from specific parts, tags and scripts in the webpage source code. Besides, ten features were extracted from the indicators of webpage URLs. This high-dimensional set of features will be refined later to a subset of selected features using several feature selection methods as it will be presented in the next subsection.

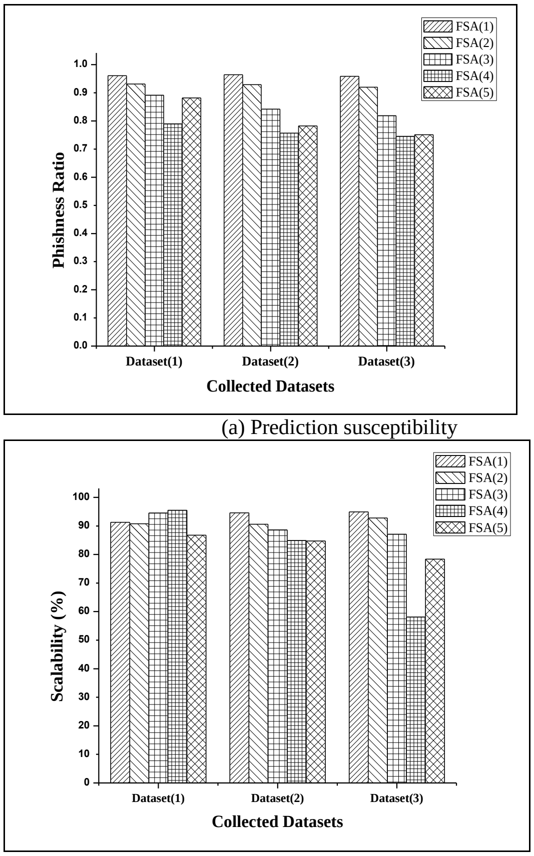
B. Comparison Across Feature Selection Methods

In this section, all the details and discussions of the first empirical test and the related findings are presented. The test was conducted on four *feature selection algorithms* (FSAs); namely *CBF*, *WFS*, χ^2, and *IG;* which had been previously adopted in the surveyed works. Besides, the *mRMR* feature selection method was also involved in the comparison to qualify if it could be recommended as an alternative FSA for the problems at hand (i.e. features' redundancy and irrelevance). Among its competitors those mentioned in *Table I*, *mRMR* discards redundant and irrelevant features in parallel and yields a selective subset of the most relevant and least redundant features together in a compact combination. Hence, both test and comparison were achieved in the presence of three datasets with different sizes and collections of phish and legitimate instances, as presented in *Fig. 3* and *Fig. 4.*

In this comparison, all the tested FSAs were practically appraised on prediction susceptibility (*Fig. 3(a)*), and scalability (*Fig. 3(b)*), goodness (*Fig. 4(a)*), stability (*Fig. 4(b)*), and similarity (*Fig. 4(c)*) to show their variations and likelihood. In *Fig. 3* and *Fig. 4*, FSAs 1, 2, 3, 4, and 5 are referring to *mRMR, CBF, WFS,* and $\chi2$and *IG* respectively.

From *Fig. 3* and *Fig. 4* , the overall results are very encouraging towards deploying all the selective hybrid features as predictive ones on phishing websites. The only difference is the variation of their compactness by using different FSAs. Findings of this test are summarized as follows:

- In *Fig. 3(a),* the evaluation and comparison of their prediction susceptibilities were done by using the measure of *Phishness Ratio* (*Table IV.*). *Phishness Ratio scores* showed that the selected feature subsets chosen by using FSAs 1, 2 and 3 (i.e. *mRMR, CBF,* and *WFS*) reached the highest peak among their competitors over all datasets; whereas the feature subsets of FSAs 4 and 5 (i.e. χ^2 and *IG*) produced the lowest peaks. Such findings point out the significance of features' mutual information for selecting the best feature subsets. More importantly, they demonstrate that the discarded redundant and irrelevant features were least predictive features among the others. And such generated compactness of most relevant and least redundant features increases their *Phishness Ratio*. Further, this test pointed out that *mRMR* criterion can be considered as a promoting technique to improve the overall discriminating behavior of the classification model websites. FSA 1 (i.e. *mRMR*) reduced both redundant and noisy features that are the prime objective of feature selection.



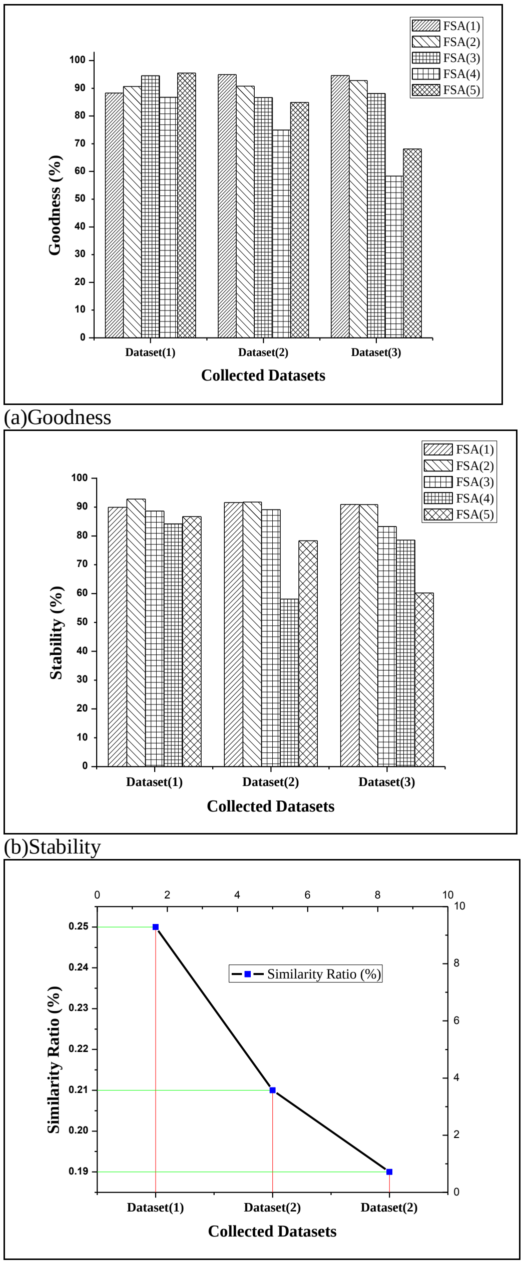
(a) Prediction susceptibility

(b)Scalability

Fig. 3. Illustration of empirical test across four five feature selection methods. Each of FSA 1, 2, 3, 4, and 5 refers to mRMR, CBF, WFS, χ^2, and IG respectively

- *Fig. 3(b)* portrays the outcomes of scalability comparison. It shows that the feature subset chosen by using FSA 1 (i.e. *mRMR*) could successfully rise the score of prediction from the typical case to the best one over datasets having different sizes. This, in turn, restates that *mRMR* can be considered as the most scalable FSA among the others because it could preserve its prediction rate as close to the best case as possible.

- *Fig. 4(a)* qualified the goodness of the selected subsets over the three different datasets. It is clearly shown that FSA 1 (i.e. *mRMR*) still preserves the best case of goodness (i.e. quality) among the others despite of the volume variations of the utilized test-bed. But both of FSAs 4 and 5 (i.e. χ^2 and *IG*) have the worst case of quality among the others. This implies that the significance of reducing feature set's dimensionality, and removing both redundant and noisy features to define the best features subset. Indeed, such feature subset will help the classification model to well perform over all datasets. More interestingly, such feature subset is needed to effectively detect phishing websites in realistic applications.



(a)Goodness

(b)Stability

(c)Similarity

Fig. 4. Illustration of empirical test across five feature selection methods; where: FSAs 1, 2, 3, 4, and 5 refer to mRMR, CBF, WFS, χ^2, and IG respectively

- *Fig. 4(b)* outlines how the feature subsets chosen by FSAs 1and 2 (i.e. *mRMR* and *CBF*) are notably more stable over all datasets than their competitors. Further, it emphasizes the significance of the inter-dependencies between the features in the same chosen feature subset. Features chosen on their inter-dependencies can compose a stable subset under different detection scenarios and datasets. In contrast, those subsets chosen with respect to the topmost ranking of their constituents like FSA 5 (i.e. *IG*) may vary in their discriminating power against vast dataset and different detection approach.

- In the context of overall outputs' similarity (*Fig. 4(c)*), it can be observed that FSAs' outputs are notably dissimilar over all the datasets. The reported similarity scores are lower than (0.3) which point out that the selected subsets overlap partially and they are complementary to each other's. Interestingly, such dissimilarity implies that feature subset composed of hybrid and diversely predictive features could be a promising avenue to improve the classification performance. Moreover, FSAs produce dissimilar feature subsets can be effectively integrated and exploited for a specific phishing detection approach. Despite this, it is clearly observed that the optimal feature subset chosen by specific FSA, it may be considered as sub-optimal choice regarding to another FSA. Hence, both likelihood and difference of FSAs outputs are crucial issue in a machine learning based detection approaches.

- Based on the overall results, we obtained a useful insight into the crucial importance of feature selection method for the problem domain at hands. This, in turn, enables us to improve the detection performance in the context of using as few, predictive and robust features as possible. In general, looking at the aforesaid test and its overall findings highlights the significance of selective feature subset in terms of prediction susceptibility, scalability, goodness, and stability. In particular, feature subset chosen by FSA 1 (i.e. *mRMR*) always has the first best scores in terms of the aforesaid perspectives among the others. Whilst, FSAs 2 and 3 (i.e. *CBF* and *WFS*) reveal the second and third best cases among the others. Contrarily, both FSAs 4 and 5 (i.e. χ^2 and *IG*) yield the worst cases across all the aforesaid perspectives.

In summary, this empirical test restates that several selection methods reach a quite bit similar peaks of prediction susceptibility and robustness. Therefore, they can be considered as the baseline methods for feature selection in phishing website detection. More importantly, if the feature selection method is carefully chosen, i.e. on the basis of its prediction susceptibility and robustness; the performance of the classification model could be highly improved with low latency and errors. However, there is still no exact answer for the perfect FSA among all the tested ones unless they assessed in terms of detection accuracy, specificity and sensitivity across several classification models and different datasets. This issue will be considered in the next subsection.

C. Comparison Across Classification Models

Herewith, we turn to qualify how the aforesaid selective subsets of features can shift detection accuracy, specificity and sensitivity of the classification model to the best rates as possible. The qualification is determined through two comparisons. First, the outputs obtained from the previously tested FSAs are compared on detection accuracy, detection sensitivity and specificity over training and testing datasets dedicated for this purpose. To accomplish the performance test and get findings for comparison, a specific machine learning classifier was applied; namely, *C4.5* as can be seen in

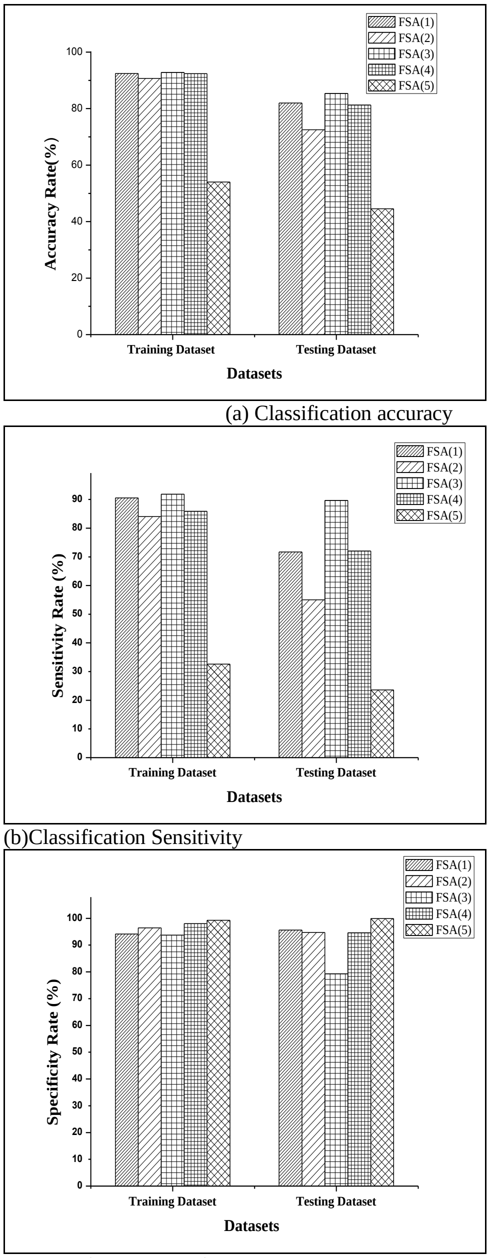
Fig. 5. Meanwhile, several supportive metrics are deployed for the performance evaluation as presented in *Table V*.

To qualify the discriminating behavior, four machine learning classifiers are involved in the second comparison. Those classifiers are described with their related calculations in *Table VI*. Such classifiers are chosen because of their wide use in the literature of phishing detection. Consequently, this comparison highlights how the best selective feature subset could classify phishing websites not only across different datasets (i.e. training and testing datasets) but also across different classification models as illustrated in *Fig. 6*.

Both comparisons are applied over two datasets: training and testing datasets that generated from a collection of phishing and legitimate webpages specifically aggregated for this purpose. The datasets are generated through extracting the features space from the aggregated webpages (i.e. data pre-processing) and dividing it into a training dataset (70% of the main dataset) and a testing dataset (30% of the main dataset).

TABLE V. PERFORMANCE EVALUATION MEASURES [1, 3]

Metrics	Evaluation Criterion	Mathematical Formula
TP	True Positive indicates the rate of correctly classified phishing instances.	$\frac{N_{P\to P}}{(N_{P\to P}+N_{P\to L})}$ (12)
FP	False Positive refers to the rate of wrongly classified legitimate instance s as phishing ones.	$\frac{N_{L\to P}}{(N_{L\to L}+N_{L\to P})}$ (13)
TN	True Negative refers to the rate of correctly identified legitimate instances.	$\frac{N_{L\to L}}{(N_{L\to L}+N_{L\to P})}$ (14)
FN	False Negative indicates the wrongly labeled phishing instances as legitimate ones.	$\frac{N_{P\to L}}{(N_{P\to P}+N_{P\to L})}$ (15)
Specificity	The percentage of correctly positive predictions	$\frac{\|TP\|}{\|TP\|+\|FP\|}$ (16)
Sensitivity	It refers to the percentage of correctly predicted positive instances (TPs).	$\frac{\|TP\|}{\|TP\|+\|FN\|}$ (17)
Accuracy	It indicates the overall rate of correctly detected phishing and legitimate instances (the rate of correct predictions).	$\frac{\|TP\|+\|TN\|}{\|TP\|+\|TN\|+\|FP\|+\|FN\|}$ (18)
Where: $N_{P\to P,}$ $N_{L\to P,}$ $N_{P\to L,}$ $N_{L\to L}$ denote the number of correctly labeled phishing instances, the number of wrongly labeled legitimate instances, the number of phishing instances that are incorrectly recognized as legitimate, and the number of legitimate instances that are identified correctly as legitimate respectively [1, 3].		



(a) Classification accuracy

(b)Classification Sensitivity

(c) Classification specificity

Fig. 5. Outcomes on classification performance with the aid of C4.5 classifier and all tested feature selection methods. Each of FSAs 1, 2, 3, 4, and 5 refers to mRMR, CBF, WFS, χ^2, and IG respectively

TABLE VI. EXAMPLS OF MACHIENE LEARNING CLASSIFIERS PREVIOUSLY ADOPTED IN PHISHING DETECTION [20, 40-44]

Machine Learning Classifier	Description	Related Calculation (s)
C4.5	It is a Decision Tree hypothesis that depends on a tree structure to construct a classification model. Its nodes represent features, its branches denote the features values whereas the leaf nodes denoting the final class decision.	The final decision of an instance to be classified relies on tracing the path of nodes and their branches to the terminating leaf nodes.
Decision Tree (DT)	It models the data with a rooted tree that contains: nodes, edges and leaves. Nodes are labeled corresponding to features, edges are labeled with the feature values and leaves are labeled with classes.	Instances of unknown class are classified by ordering them according to their feature values in the rooted tree such that features are denoted by nodes and their values are represented by branches that the node assumes. The classification of unknown instance is started at the root node and then passed through the tree. The test at each node along the path is applied to the sorted feature values that determine the next edge until ending up at the leaf nodes. The label of the ended up leaf node is the final decision of classification.
Naïve Bayes (NB)	A probabilistic classifier with assumption of conditionally independent attributes of each other given class of instances.	$P(C\|X) = P(C\|x_1, \dots, x_n) = \frac{P(C)P(x_1,\dots,x_n\|C)}{P(x_1,\dots,x_n)}$ (19) Where X is a given sample with a vector of *n* features$(x_1, \dots, x_n)$, C is the class label that the classifier seeks for maximizing the likelihood.
Support Vector Machine (SVM)	It is an optimistic separating hyper-plane that maximizes the margin between closest points of two classes to estimate the decision function.	$min\,\frac{1}{2}w^T w + C\sum_i \xi_i$ (20) Subject to: $y_i\big((w^T \cdot x_i) + b\big) \geq 1 - \xi_i$, $\xi \geq 0, i = 1, 2, \dots, m,$ (21) $max\sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$ (22) Subject to:$0 \leq \alpha_i \leq C, i = 1, 2, \dots, m\ and\ \sum_{i=1} \alpha_i y_i = 0.$ (23) Where x_i is M-dimensional data vector $x_i \in R^m$ with samples belong to either one of two classes labeled as $y \in \{-1, +1\}$that it is separated by a hyper-plane of $(w \cdot x) + b = 0$. $\alpha_i$denotes the Lagrange multipliers for each vector in the training dataset and it is used to transform the original input space to higher in dimension space.
Transductive Support Vector Machine (TSVM)	It separates the positive and negative samples included in the training dataset with a maximal margin by using SVM hyper-plane. It outperforms SVM with good generalization accuracy.	Minimize over $(y_1^*, \dots, y_k^*, w, b, \xi_1, \dots, \xi_n, \xi_1^*, \dots, \xi_k^*)$, $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_1 + C'\sum_{j=1}^{k} \xi_j^*$ (24) Subject to: $\forall_{i=1}^{n}: y_i[wv_i + b] \geq 1 - \xi_i$, $\forall_{j=1}^{k}: y_i[wv_i^* + b] \geq 1 - \xi_j^*, \forall_{i=1}^{n}: \xi \geq 0$, and $\forall_{j=1}^{k}: \xi_j^* \geq 0.$ (25) Where x_i is an *m*-dimensional vector such that$x_i \in R^m$ with independent labeled samples belong to either one of two classes labeled as$y \in \{-1, +1\}$, ξ and ξ^* are the slack variables of training and testing datasets, respectively. C and C' denote the influencing parameters determined by the user. The effect term of the jth unlabeled sample is denoted by $c^*\xi_j^*$.

Regarding *Tables V and VI* as well as the statistics plotted in *Fig. 5* and *Fig. 6*, the following standpoints are inferred:

- The significant differences between classification models assisted by the tested FSAs (*Fig. 5*) point out the major or minor contribution that the assisted feature selection method can provide. Variations in accuracy, sensitivity, and specificity demonstrate that not all the tested feature selection method yield promising outcomes on phish website detection. This is because of (i) variations on specifics and evaluation criteria of FSAs themselves, (ii) the chosen features themselves due to their varied prediction susceptibilities and robustness, (iii) the inter-dependency of detection performance on the deployed classification model itself, (iv) the type of exploited features (i.e. webpage's URL and /or webpage's content) and (v) the dimension of the selected feature subset (i.e. the number of features included in the selected subset).
- Consequently, different outcomes of performance test (*Fig. 5*) show that certain classification model may sensibly being influenced by the training and testing datasets, and the suitability of machine learning classifier as well as the chosen feature selection method. This implies that the diversity and pre-processing of the collected dataset likely influence the overall classification performance because the dataset may encompass imbalanced data. More precisely, the imbalanced data indicate the divergent abundance of features corresponding to the classes of phishing or legitimate over the collected test-bed. Since the collected test-bed is quite bit different in dataset size and it consists of a dozen of labelled and unlabeled instances having a variety of features (i.e. hybridity), and a heterogeneity of features values. Therefore, k-fold validation and chronological assessment must be attained to come up with such diversity.

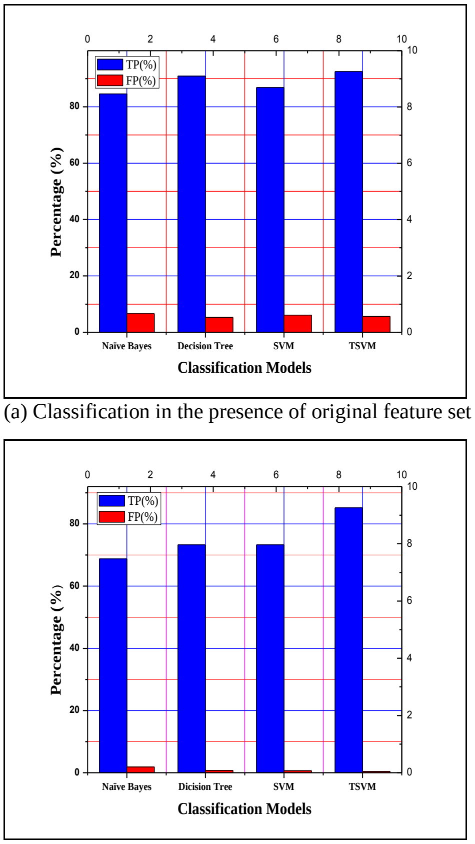
- The classification performance is likely to be influenced by the set of many features (*Fig. 6 (a)*). For instance, 58 extracted hybrid features may encompass irrelevance, redundancy and noisy data; therefore, eliminating the worst features and selecting the best ones (i.e. the most representative ones) are important inductive factors for well-performed classification as can be recognized in *Fig. 6(b)*.

- Also, the feature set's dimensionality is an important factor for the classification performance (*Fig. 6(a) and Fig. 6(b)*). As more features are being processed as more computational cost is being consumed. Moreover, the feature set's dimensionality interacts with the dataset's dimensionality.

- Selected feature subset chosen by the *mRMR* promotes the overall performance of classification models. Classification models assisted by *mRMR* outperform those baseline models in terms of classification accuracy and error rates (*Fig. 6(a) and Fig. 6(b)*).

V. Conclusions and Future Work

In the light of selecting a minimal and effective feature subset for well-performed phish website detection technique, this paper critically and practically appraised the exploitation of the feature selection via classification-based techniques. In this appraisal, those techniques assisted by machine learning classifiers and feature selection methods were involved, as well as a review of prior works with their related issues.



(a) Classification in the presence of original feature set

(b)Classification in the presence of selective feature subset chosen by mRMR

Fig. 6. Outcomes on classification performance across different classifiers

Further, empirical tests are conducted over 58 new hybrid features, five different datasets and five different classification models. Promoting measures are introduced to assess the outcomes of applied feature selection methods and then qualify the most suitable one among them for the problem at hands. Deeper understanding to their effects and significant gains on their outcomes' prediction susceptibility, scalability, goodness, stability and similarity are obtained respectively. Moreover, feature selection outcomes are compared on how they can notably improve the overall classification performance towards finding an optimal anti-phishing solution.

As a result, the findings displayed that some feature selection methods significantly outperformed their competitors by exhibiting better robustness, prediction, and performance. Between, other methods diverted from the best and the worst cases in relation to the aforesaid quantified factors. This was caused by the variations in dataset sizes and their constituent instances, the compactness of the chosen features and the features themselves, the evaluation criteria of the selected methods, and the discriminating behavior of the applied classifiers on training and testing instances. Moreover, the empirical tests addressed that the appropriately chosen set of features outperformed the original set of extracted features and/or the individual features themselves with least latency. However, the notably powerful selection method (i.e. *mRMR*) failed to provide an ideal subset of features; it could only produce as minimal and effective feature subset as possible. Nonetheless, *mRMR* could deal with the problematic features of redundancy and irrelevance at once. However, it is worthy to mention that no precise feature selection method existed in this study to cope with all the classification models. Hence, the forthcoming work will quantify feature selection outcomes concerning the processing time and misclassification costs. With that, more classification models will be involved in a remedial framework for feature selection towards rational phish website detection.

References

[1] M. Khonji, Y. , Iraqi, A. and Jones, "Phishing detection: a literature survey", Communications Surveys & Tutorials, IEEE. , (15), 2091-2121, 2013.

[2] S. Purkait, "Phishing counter measures and their effectiveness–literature review", Information Management & Computer Security, (20), 382-420, 2012.

[3] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques", Communications Surveys & Tutorials, IEEE., 15(4).2070-2090, 2013.

[4] R. Islam, and J. Abawajy, " A multi-tier phishing detection and filtering approach", Journal of Network and Computer Applications, (36).324-335, 2013.

[5] G. Ramesh, I. Krishnamurthi, and K. Kumar, "An efficacious method for detecting phishing webpages through target domain identification", Decision Support Systems, (61).12-22, 2014.

[6] P. Barraclough, M. Hossain,M. Tahir, G. Sexton, and N. Aslam, "Intelligent phishing detection and protection scheme for online transactions", Expert Systems with Applications, (40).4697-4706, 2013..

[7] H. Shahriar, and M. Zulkernine, "Trustworthiness testing of phishing websites: A behavior model-based approach", Future Generation Computer Systems, (28).1258-1271, 2012.

[8] M. Bhati, and R. Khan, "Prevention Approach of Phishing on Different Websites", International Journal of Engineering and Technology, (2), 2012.

[9] M. He, S.-J. Horng, P. Fan, M. M. Khan, R.-S. Run, and J.-L. Lai, " An efficient phishing webpage detector", Expert Systems with Applications, (38), 12018-12027, 2011.

[10] W. Han, Y. Cao, E. Bertino, and J. Yong, "Using automated individual white-list to protect web digital identities", Expert Systems with Applications, (39). 11861-11869, 2012.

[11] S. Gastellier-Prevost, G. G. Granadillo, and M. Laurent, "Decisive heuristics to differentiate legitimate from phishing sites", 2011 Conference on Network and Information Systems Security (SAR-SSI), 1-9, 2011.

[12] H. Wang, B. Zhu, and C. Wang, "A Method of Detecting Phishing Web Pages Based on Feature Vectors Matching", Journal of Information and Computational Systems, (9).4229-4235, 2012.

[13] Y. Chen, Y. Li, X. Q. Cheng, and L. Guo, "Survey and taxonomy of feature selection algorithms in intrusion detection system", In Information Security and Cryptology, Springer Berlin Heidelberg, 153-167, 2006.

[14] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research", ASU feature selection repository, 2010.

[15] P. Likarish, E. Jung, D. Dunbar, T. E. Hansen, and J. P. Hourcade, "B-apt: Bayesian anti-phishing toolbar", IEEE International Conference on Communications, ICC'08, 1745-1749, 2008.

[16] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages", NDSS, 2010.

[17] A. Bergholz, J. De Beer, S. Glahn, M.-F. Moens, G. Paaß, and S. Strobel, "New filtering approaches for phishing email. Journal of computer security", 18(1), 7-35, 2010.

[18] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: a feature-rich machine learning framework for detecting phishing web sites", ACM Transactions on Information and System Security (TISSEC), (14). 21, 2011.

[19] H. Zhang, G. Liu, T. W. Chow, and W. Liu, "Textual and visual content-based anti-phishing: a Bayesian approach", IEEE Transactions on Neural Networks, 22(10), 1532-1546, 2011.

[20] Y. Li, R. Xiao, J. Feng, and L. Zhao, "A semi-supervised learning approach for detection of phishing webpages", Optik-International Journal for Light and Electron Optics, (124), 6027-6033, 2013.

[21] H. Kordestani and M. Shajari, "An entice resistant automatic phishing detection", 2013 5th Conference on Information and Knowledge Technology (IKT), 134-139, 2013.

[22] R. Gowtham, and I. Krishnamurthi, "A comprehensive and efficacious architecture for detecting phishing webpages", Computers & Security, (40), 23-37, 2014.

[23] Y. Pan, and X. Ding, "Anomaly based web phishing page detection", In 22nd Annual 2006 Computer Security Applications Conference, (ACSAC'06), IEEE., 2006.

[24] L. Ma, B. Ofoghi, P. Watters, and S. Brown, "Detecting phishing emails using hybrid features", Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC'09), IEEE., 493-497, 2009.

[25] M. Khonji, A. Jones, and Y. Iraqi, "A study of feature subset evaluators and feature subset searching methods for phishing classification", In Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, ACM, 135-144, 2011.

[26] R. B. Basnet, A. H. Sung, and Q. Liu, "Feature selection for improved phishing detection", In Advanced Research in Applied Artificial Intelligence, Springer Berlin Heidelberg, 252-261, 2012.

[27] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites", Information & Management, 51(7), 845-853, 2014.

[28] I. R. A. Hamid, and J. H. Abawajy, "An approach for profiling phishing activities", Computers & Security, (45), 27-41, 2014.

[29] C. M. Chen, H. M. Lee, and Y. J. Chang, "Two novel feature selection approaches for web page classification", Expert systems with Applications, 36(1), 260-272, 2009.

[30] L. Yu, and H. Liu, "Efficient feature selection via analysis of relevance and redundancy", The Journal of Machine Learning Research, 5, 1205-1224, 2004.

[31] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, "Toward an efficient and scalable feature selection approach for internet traffic classification", Computer Networks, 57(9), 2040-2057, 2013.

[32] Z. He, and W. Yu, "Stable feature selection for biomarker discovery", Computational Biology and Chemistry, 34(4), 215-225, 2010.

[33] N. Dessì, and B. Pes, "Similarity of feature selection methods: An empirical study across data intensive classification tasks", Expert Systems with Applications, 42(10), 4632-4642, 2015.

[34] M. Khonji, Y. Iraqi, and A. Jones, "Lexical URL analysis for discriminating phishing and legitimate websites", Anti-Abuse and Spam Conference Proceedings of the 8th Annual Collaboration, Electronic Messaging, ACM, 2011.

[35] S. Lee, Y. T. Park, and B. J. Auriol, "A novel feature selection method based on normalized mutual information", Applied Intelligence, 37(1), 100-120, 2012.

[36] S. Tabakhi, and P. Moradi, "Relevance–redundancy feature selection based on ant colony optimization", Pattern Recognition, 48(9), 2798-2811, 2015.

[37] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238, 2009.

[38] E. Uzun, H. V. Agun, and T. A. Yerlikaya, "A hybrid approach for extracting informative content from web pages", Information Processing & Management, (49), 928-944, 2013.

[39] L. Fu, Y. Meng, Y. Xia, and H. Yu, "Web content extraction based on webpage layout analysis", 2010 Second International Conference on Information Technology and Computer Science (ITCS), 40-43, 2010.

[40] A. Ben-Hur, and J. Weston, "A user's guide to support vector machines", Data mining techniques for the life sciences, pp. 223-239, Springer, 2010.

[41] G. Kumar, K. Kumar, and M. Sachdeva, "The use of artificial intelligence based techniques for intrusion detection: a review", Artificial Intelligence Review, 34(4), 369-387, 2010.

[42] D. Miyamoto, H. Hazeyama, and Y. Kadobayashi, "An evaluation of machine learning-based methods for detection of phishing sites", Advances in Neuro-Information Processing, pp. 539-546, Springer, 2009.

[43] H. Patel, and J. Sarvakar, "Analysis of data mining algorithm in intrusion detection", International Journal of Emerging Technology and Advanced Engineering (IJETAE), 1(1), 2011.

[44] C.-F Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review", Expert Systems with Applications, 36(10), 11994-12000, 2009.

APPENDIX I. THE ORIGINAL FEATURE SET CONSISTS OF 58 HYBRID FEATURES

Webpage Content Features					
Index	***Feature***	***Type***	***Index***	***Feature***	***Type***
F1	Number of Scripting.FileSystemObj	Continuous	F24	Number <input> in java scripts	Continuous
F2	Number of Excel.Application	Continuous	F25	JavaScript scripts length	Continuous
F3	Presence of WScript.shell	Discrete	F26	Number of functions’ calls in java scripts	Continuous
F4	Presence of Adodb.Stream	Discrete	F27	Number of script lines in java scripts	Continuous
	Presence of Microsoft.XMLDOM	Discrete	F28	Script line length in java scripts	Continuous
	Number of <embed>	Continuous	F29	Existence of long variable names in java scripts	Discrete
	Number of <applet>	Continuous	F30	Existence of long function names in java scripts	Discrete
	Number of Word.Application	Continuous	F31	Number of fromCharCode()	Continuous
	link length in <embed>	Continuous	F32	Number attachEvent()	Continuous
	Number of <iframe>	Continuous	F33	Number of eval()	Continuous
	Number of <frame>	Continuous	F34	Number of escap()	Continuous
	Out-of-place tags	Discrete	F35	Number of dispacthEvent()	Continuous
	Number of <form>	Continuous	F36	Number of SetTimeout()	Continuous
	Number <input>	Continuous	F37	Number of exec()	Continuous
	Number of MSXML2.XMLHTTP	Continuous	F38	Number of pop()	Continuous
	Frequent <head>, <title>, <body>	Discrete	F39	Number of replaceNode()	Continuous
	<meta index.php?Sp1=>	Discrete	F40	Number of onerror()	Continuous
	“Codebase” attribute in <object>	Discrete	F41	Number of onload()	Continuous
	“Codebase” attribute in <applet>	Discrete	F42	Number of onunload()	Continuous
	“href” attribute of <link>	Discrete	F43	Number of <script>	Continuous
	Number of void links in <form>	Continuous	F44	frequent<div onClick=window.open()”>	Discrete
	Number of out links	Continuous	F47	Number of onerror()in javascripts	Continuous
	Number of <form> in java scripts	Continuous	F48	Number of SetInterval()	Continuous
URL Features					
Index	***Feature***	***Type***	***Index***	***Feature***	***Type***
	Multiple TLD	Discrete	F54	Typos in Base name	Discrete
	Brandname in hostname	Discrete	F55	Long domain name	Discrete
	Special symbols in URL	Discrete	F56	Misleading subdomain	Discrete
	Coded URL	Discrete	F57	Number of dots in URL	Continuous
	IP address instead of domain name	Discrete	F58	Path domain length	Continuous

Computer Science Approach to Information-Like Artifacts as Exemplified by Memes

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

***Abstract*—Providing information can be expanded to include systems that deliver *information-like* artifacts. They provide such "things" as advertisements, propaganda pieces, and meme artifacts. Memes are the subject of extensive intellectual debate in science and popular culture because it is claimed that parallels can be drawn between theories of cultural evolution manifested in memes, and theories of biological evolution. Memes are described as self-reproducing mental structures, intangible entities transmitted from mind to mind, verbally or by repeated actions and/or imitation. The problem is that researchers describe memes in terms of English-language text or ad hoc diagrams. This paper considers the problem that the field of memetics lacks a uniform language for examining diverse conceptualizations of memes. The paper presents a unifying diagrammatic representation used in computer science, in which all types of "claimed" memes can be expressed and their general characteristics observed. Several examples from the literature on memes are recast in terms of this representation. The results point to the capability of the proposed depiction to express various types of memes.**

Keywords—information; cultural evolution; mental structures; memes; memetics; conceptual modeling; diagrammatic representation

I. Introduction

Information Systems attempts to provide the business client with information that maximizes its effectiveness [1]. Going further, it is possible to remove the restriction of providing information, to include systems that provide clientele with information-like artifacts that maximize their effectiveness. They include systems that provide such "things" as advertisements, propaganda pieces, and meme artifacts. Memes, the topic of this paper, are of special importance in this context, expressed as follows:

> We are potentially facing what is termed a General Purpose Technology or Disruptive Innovation. If this prediction proves right, the world will be ripe for a new version of James Brown's "It's a Man's Man's Man's World" to be titled "*It's a Memes' Memes' Memes' World*" [2]. (Italics added)

This paper tries to strengthen the notion of *meme* by proposing a schematic representation of transmission of memes. Memes (cultural artifacts) are interesting "flow things" because of their complex conceptual relationships with "life things" or genes.

> Few scientific terms introduced into scientific and popular vernacular have enjoyed the impact on intellectual debate as has the term "meme."… The meme concept has generated recent excitement precisely because it seems to offer hope of providing something that other theories of social and semiotic processes have not succeeded in providing. [3].

Memes are usually described as *units of cultural flow* that may change in terms of host state (temporal/special), population (growth through replication), and variety (differences). The following paragraphs highlight some of the literature pertaining to memetics: "the theoretical and empirical science that studies the replication, spread and evolution of memes" [4]. The focus here is on more description of the notion of meme, the life cycles of memes, and modeling of memes. Modeling in this context refers to development of an abstract representation of a meme life cycle through diagrammatic representation of its flow from birth to its destination, including its effects on its hosts.

A. *About Memes*

According to Dawkins [5],

> We need a name … that conveys the idea of a unit of cultural transmission, … I want a monosyllable that sounds a bit like "gene"… it could alternatively be thought of as being related to "memory"… Just as genes propagate themselves in the gene pool by leaping from body to body via sperms or eggs, so memes propagate themselves in the meme pool by leaping from brain to brain via a process which, in the broad sense, can be called imitation. (p. 192)

Memes are also described as self-reproducing mental structures, intangible entities, transmitted from mind to mind, either verbally, with actions, music, graffiti, leaflets, television broadcasts, or by repeated actions and/or imitation. Hence, they have been viewed as "a base for constructing an illustrative model of social and cultural behavior raising questions about how memes are generated, received, transmitted, replicated, and re-transmitted" [6]. Memes "influence ideas, ideas influence and form beliefs ..., eventually producing actions" [6].

The replication and transmission of memes in culture have been described in analogy to biological genes [7]. A combined group of memes is called a memeplex. A central concept in the meme-gene analogy is replication. A replicator is first defined as a unit that has largely the same structure before and after a process of copying or replication is completed. Some errors may be made in the process, but too many errors are not allowed for a copying process to count as a replication process [8].

The second essential characteristic [of a meme] is that it is part of a lineage. Genes are called replicators because they are copied over and over. Over and over means that a replicator is first copied from a mould, then in a next replication event it is used as a mould itself, from which a new copy is produced, that again will function as a mould in the next event, and so on [8].

The media by which a meme flows (e.g., journals, radio stations) affects its life cycle. Computer networks currently play a major role as a meme environment. An Internet meme is a meme that spreads online, e.g., via online social networks. It refers to the propagation of content items such as jokes, videos, and websites from one person to another via the Internet [9].

B. *Meme Life Cycle*

According to Heylighen [10], memes go through a four-stage life cycle: assimilation, transmission, expression, and retention. Assimilation happens when a meme arrives at a new host and is accepted and processed. At this stage, the “meme is very vulnerable to misunderstanding” [11]. Expression is when “a meme must emerge from its storage as memory pattern and enter into a physical shape that can be perceived by others” [10]. Transmission is when the meme is presented through some form of media. Retention is when the host “memorizes” the meme.

For Bjarneskans, Grønnevik, and Sandberg [12], the meme life cycle is similar to that of parasites. It includes transmission, decoding, infection, and coding phases. Encoding of the meme happens in the transmission phase, when the meme is coded in a vector such as spoken message, text, image, e-mail, or observed behavior. It is “some kind of information-carrying medium.”

When a potential host decodes the meme … the meme may become active and infect the person, who becomes a new host (the infection phase). At some point the meme is encoded in a suitable vector (not necessarily the same medium it was originally decoded from) and can be spread to infect new hosts [12].

C. *Modeling and Representation*

The brief, general description of memes and their life cycle presented above establishes this paper’s topic and its associated discourse. This subsection closes in on the center of interest: modeling, and in particular producing a conceptual representation of the flow of memes and learning the significance of that flow.

Much of the published research draws parallels between theories of cultural evolution and those of biological evolution [13]. A typical dictionary definition of *evolution* describes it as a process of continuous *change* of something into a different and usually more complex or new form. Change is a fundamental notion with a long history. Heraclitus [14], a pre-Socratic Greek philosopher, observed that everything changes and nothing remains still [3]. He described this change as a *flow* of things, declaring that "everything flows." Flow can be observed everywhere, especially in nature and in social ontology, e.g., flows of water, electricity, gas, money, materials, memes, and so on.

Such a view aims at capturing a real-world process, specified in text or diagram form or written in a formal language. “The resulting depiction expresses a process model, a process template, process metadata,… A process definition normally comprises a number of discrete activity steps, with associated computer and/or human operations and rules governing the progression of the process through the various activity steps” [15]. In this perspective, modeling provides a framework for organizing and enabling future exploration [16]. In the context of the current topic, it is used to observe the path of memes moving through hosts in order to develop a broader theoretical basis for examining the changes that occur in memes over time [16].

For example, Husted [16] proposes a multilevel model for cultural meme transference based on the behavior of the ocean ecosystem. “The ocean works as a model because organic materials move through the ocean in the same way that cultural objects move through society” [16]. Communities of memes can be compared with a coral reef within a larger ocean ecosystem; similarly, the Internet is an adaptive entity consisting of many cooperative communities of organisms [16]. Coral reef production of biodiversity is analogous to the forum cultures that produce the largest number of memes on the Internet.

A coral reef represents a meme-producing core society. Different inhabitants are specified in the model, e.g., lurking users who do not participate in the communities they are observing, along with core memes that attack inhabitants of other levels.

D. *Problem and Solution*

This paper addresses the problem of *lack of a uniform model for examining diverse conceptualizations of memes*. For example, a question answered in the context of this problem: How to represent such diverse things as *graffiti* and *strategy*, which have been claimed to be memes, in a uniform way so their common characteristics (e.g., evolution, spreading, multiplicity, …) can be compared? The paper presents a unifying conceptual representation in which all types of “claimed” memes can be expressed and their general characteristics observed.

For the sake of a self-contained paper, we next review the model on which our representation of the flow of memes will be built. After that, in the section titled “Schematizing Memes,” a meme is described as a special type of “flow thing” by use of schemata (patterns of streams of flow) based on this model.

The section models changes in a meme through flow across hosts by drawing schemata of flows, including five stages in the life cycles of each host. The fourth section, titled “Applying FM to Sample Meme Systems,” applies this representation of meme flow in three areas: organizational strategy, human-centered models of processes, and agents’ decisions with regard to meme adoption and transmission.

II. Flowthing Model

The Flowthing Model (FM) [17-22] represents some segment of reality as a web of interrelated *flows* that cross boundaries of intersecting and nested *spheres*. In the context of memes, every host is represented by a sphere with a hierarchical structure. Ingredients in a flow include *flowthings* (things that flow, e.g., memes), and their systems (*flowsystems*): a structure of flow comprising at most six stages (see Fig. 1). A "thing" is defined as a flowthing: what is created, released, transferred, arrived, accepted, and processed while flowing within and among *spheres*. It has a permanent identity but impermanent form, e.g., the same memes translated into different languages. A *flowsystem* constrains the trajectory of flow of flowthings. A particular flowsystem is the space/time context for *happenings* (e.g., received, released) and existence of flowthings. From the perspective of flowthings, the flowsystem is formed from six discontinuities: being *created*, being *released*, being *transferred*, being *arrived*, being *accepted*, and being *processed*.

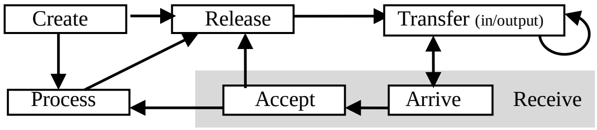


Fig. 1. Flowsystem

Flows connect six *states* (also called stages) that are exclusive for flowthings; i.e., a flowthing can be in *only one* of these six states at a time: transfer, process, creation, release, arrival, or acceptance, as shown in Fig. 1, analogous to water being in one of three states in Earth's atmosphere: solid, liquid, or gas. A *state* here is a "transmigration field" of the flowthing that is created, processed, and released, transferred, arrives, and is accepted (or is simply received, combining arrived and accepted into one state). In Fig. 1, we assume irreversibility of flow, e.g., released flowthings flow only to Transfer.

The exclusiveness of FM states (i.e., a flowthing cannot be in two states simultaneously) indicates synchronized change of the flowthing, e.g., a flowthing *cannot be changed in form and sphere simultaneously*. This is a basic systematic property of flowthings. Note the generality of the notion of flow in FM. For example, *creation of* a flowthing is a *flow* (from *nonexistence*, i.e., not currently existing in the system, to *existence*, i.e., appearance in the system).

Initialization, stopping, and continuing of flows occur through *triggering*: a control mechanism. It is the only linkage among elements in FM description besides flow and is indicated by dashed arrows. Synchronizations (e.g., join/fork) and logic notions (e.g., and/or) can be superimposed over the basic FM depiction. Note that these mechanisms can be modeled as flowsystems.

III. Schematizing Memes

The thesis of this paper is that a meme is a special type of flowthing with a distinctive schema or pattern of streams of flow. Several works have been published classifying memes as a general category. Deacon [3], in semiotics, suggests that "Memes are signs, or more accurately, sign vehicles… They are … concrete things, or events, or processes."

In FM, a general conceptualization of a meme (see Fig. 2) is similar to that of a message with a source flowsystem that creates it (circle 1 in the figure), and hosts that process and duplicate it. In Fig. 2, the curved arrow (3) indicates the flow of the meme to many other hosts (flowsystems).

The meme is received, processed, and replicated (stored – cylinder shape, circle 4) in the received state or during the process stage and may then flow to other hosts. Here, replication assumes *not creating* a new meme different from the original, which is not always the case, as will be discussed later. Fig. 2 represents more than copying and passing along, but also *affecting* (5) the meme's host.

In the Fig.2 representation, the flow of the meme to the infected hosts is not necessarily linear (from one to another); rather, it is understood as a fanned-out flow (one-to-many). Also, the meme is replicated in the Receive and Process stages to indicate the possibility of replication of the original received meme or a modified version after processing; e.g., when an idea is passed from one individual to another, not all the exact details are replicated.

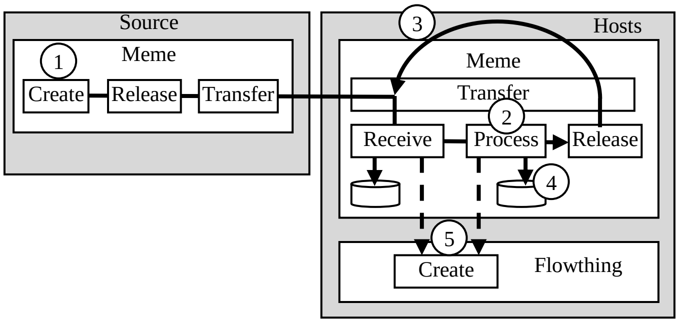


Fig. 2. A meme schema in FM

In this case the FM schema (the totality of the FM diagram) is analogous to a city map that shows all possible streets of traffic flow. Note that, as there is a *source* that creates a meme, its dissemination is analogous to the flow of a river through a basin as it forms a delta; eventually, memes dry out where flowsystems *de-create* them. This de-creation is not shown in Fig. 2 but will be shown later. Also, the lumping together of memes to form a complex meme is not shown in the figure.

Additionally, it is typically said that a meme "infection" influences a host's ideas, beliefs, behavior, etc. Hence the receiving or processing of a meme triggers (circle 5) the creation of some flowthing (e.g., an action) in the host.

Example: According to Brodie [23], a meme is a strategy, a rule of thumb for handling a situation in order to achieve some result. Typical strategy-memes in driving behavior: When you get to a traffic circle, go counter-clockwise. When you come to a red light …, If it is green…, etc. Fig. 3 shows the flow of driving-strategy memes passing through the sphere of a Driver. Fig. 4 shows an instance of activation of the strategy when a certain traffic-light meme flows to the driver.

As mentioned in the introduction, a typical dictionary definition of evolution describes it as a process of continuous change of something into a different and usually more complex or new form. In FM, this change arises from streams of flow (flows across flowsystems). In Fig. 5, a meme can be changed in any host during the processing stage, but this change does not create a new meme; however, with this continuous change it is possible that a new meme can be created along the flowstream, as shown in the right-hand flowsystem of Fig. 5. Note that this flowsystem includes the Create stage.

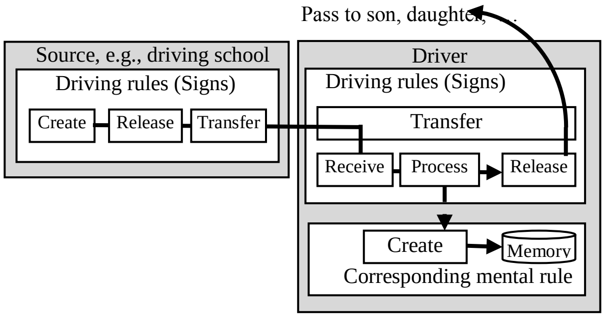


Fig. 3. Example of a meme schema in FM

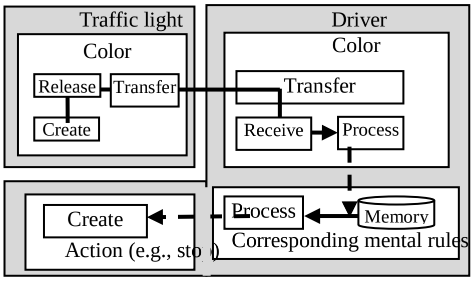


Fig. 4. Traffic light and driver's reaction

Nevertheless, it seems that other types of schemata of memes exist besides the known general form of flow of memes, as described in the next section.

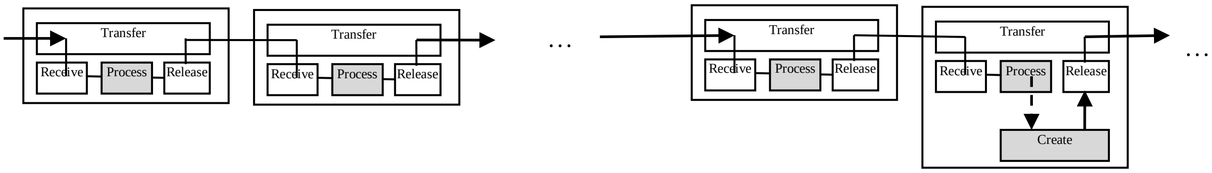


Fig. 5. Evolution of a meme

IV. Applying FM to Sample Meme Systems

This section recasts four applications of memes in terms of FM. The purpose is to demonstrate that FM can uniformly express diverse situations involving memes.

A. *Organizational Strategy*

Speel [8] developed a description of a class of memetic *evolution of an organizational strategy* to achieve goals, with the aim of showing how such evolution can be described by memetics and illustrate conceptual evolution in science. According to Speel [8], a strategy is a "program of *actions* to be taken in order to achieve agreed upon *goals*" [italics added]. It is a plan to change things to solve particular *problems*, by defining goals to be solved and viewed as a meme-complex that evolves by selective processes, including evolutionary learning. It is assumed that an organization creates a new version of its strategy through changes in goals, and actions. The actions are connected to the goals by knowledge. New knowledge can be acquired in a kind of evolutionary learning that can take place. The evolution of a strategy is a process encompassing various players involved in different selective events. For example, these players could be categorized as having high (e.g., high-level executives), middle (e.g., middle management), or low-level judgment responsibility. [8]

Fig. 6 shows an illustration of the flows and triggering involved in the organizational strategy under consideration. A strategy is developed and flows to various units of the organization to create actions that trigger feedback used to create a new version of the strategy. The strategy is a meme that flows to different organizational units and triggers them to do actions. The units also have subunits, and the strategy flows to these subunits to trigger sub-actions.

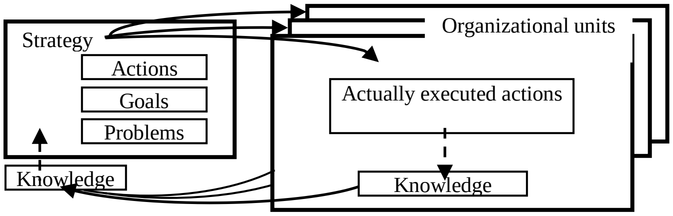


Fig. 6. Illustration of flows and triggering in an organization

It is assumed that a strategy team is responsible for developing and maintaining different versions of the strategy. In Fig. 7,

- The processing of knowledge (circle 1), problems (circle 2), goals (3), and prescribed actions (4) triggers the creation of a strategy (5).
- The processing of knowledge (6) and problems (7) triggers the creation of goals (8).
- The processing of goals (9) triggers the creation of prescribed actions

The thick vertical bar denotes a logical join. Knowledge, problems, goals, actions, and the entire strategy are flowthings that can be created, released, transferred, received, and processed. The strategy is a sphere that includes the subspheres of problems, goals, actions, and its physical self. The last flowthing (physical self) is drawn without a box and includes the stages of create (5), release, and transfer (11). The difference between the Strategy sphere and the Physical Strategy is similar to the difference between a person (as a cultural unit) and his/her physical body.

The FM representation presents a complete picture (guaranteed by continuity of flows) of all changes involved:

appearance of new flowthings, changes in stages, in spheres, ... The schematic depiction provides a unified foundation for monitoring the evolution of all memes; otherwise, the evolution is written out in English.

Fig. 7 is still based on the general life cycle of a meme discussed in Fig. 2, where a flowthing flows to a host and affects that host (executes actions). Duplication occurs in terms of making copies available to all organizational units; however, this method of transmission does not seem to align with the popular conception of memes that spread like viruses.

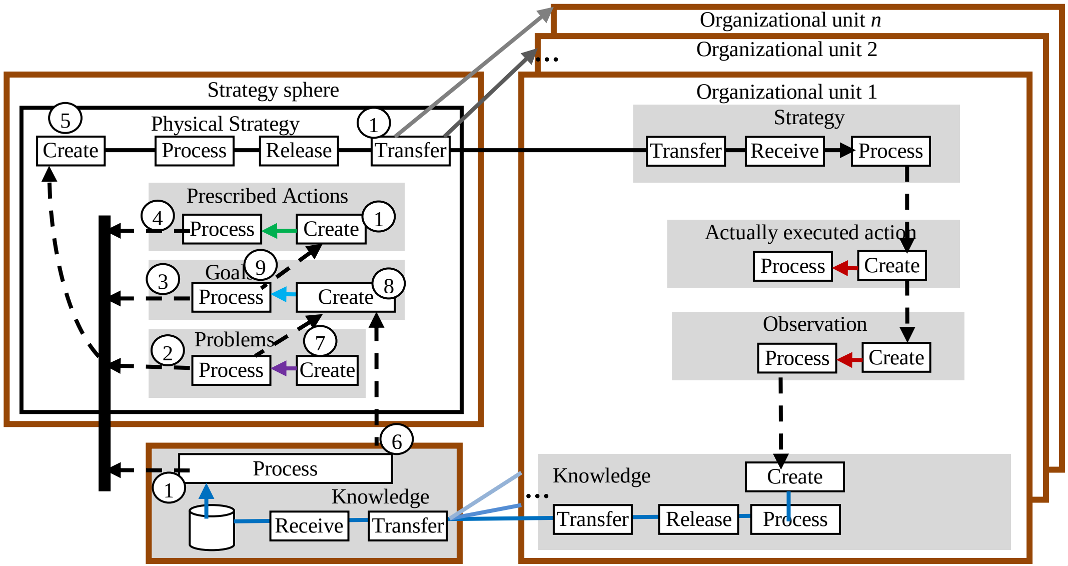


Fig. 7. FM representation of strategy flow in an organization

In terms of a meme as described by Speel [8], does strategy have the same general behavior as, say, *graffiti* in terms of multiplicity and spreading? Answering this type of question is not the aim of this paper. The paper presents a unifying language (FM diagrams) in which all types of "claimed" memes can be expressed and their general characteristics observed.

B. *Human-Centered Models of Processes*

According to Noessel [24],

Have you ever been in a design review where instead of talking about the proposed solution you spend half the time revisiting what the user is trying to accomplish in the first place? Keeping the human-centered models of the processes that lie behind your solution fresh in the minds of stakeholders (and designers) can prevent this unwanted rehashing. One way to ensure this is to create a diagram and give it qualities that make it simple enough and memorable enough so that, on a dime, you can whip out a dry-erase pen and sketch it out as a reminder…

So what about the form of the diagram? Answering this question takes us into the heady world of memes and memetics, but we'll dip just enough to come back with some meaningful attributes… *A meme can be informally defined as an information structure that replicates between human minds. Examples include ... simple business process diagrams*. [Italics added]

Noessel [24] gives an example of development of service for a company that supplies fire extinguishers. Six phases of the consumer product life cycle are identified, as follows (Fig. 8):

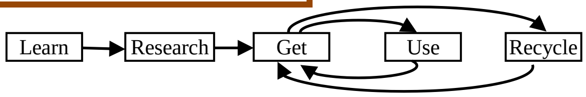


Fig. 8. Business process in a company that supplies fire extinguishers (redrawn from Noessel [24])

- Learning the need for the product
- Researching
- Acquiring the product
- If there is a fire, using the product (leading to the need for another)
- Learning when it expires (leading to the need to recycle it)
- Recycling the product when its chemicals expire (leading to need for another)

Fig. 9 shows the FM representation that corresponds to this life cycle, according to our understanding of Noessel's [24] description.

First, *learning and research artifacts* are received and processed (circles 1 and 2 in the figure). This triggers knowledge (circle 3). The thick vertical bar (4) indicates a joint operation. Note that this has been drawn as such for the sake of simplicity; however, it itself can be drawn as an FM diagram. Knowledge triggers the creation (5) of orders that, in turn, trigger (6) the production of fire extinguishers. Fire extinguishers flow to the utilization flowsystem (7), where they are categorized and processed as follows:

- Used, hence trigger new orders to produce more fire extinguishers (8)

- Expired, hence they flow to recycling (9), where they are refilled with chemicals, then flow back to be utilized (10).

Note that Order in Fig. 9 has been added to the FM description even though it was not mentioned in Noessel's (2008) original scenario, where in Graph 8 *Learn* and *Research* are followed, abruptly, by *Get*. In FM, the notion of flow ties events together in a continuity of narration: *Learn* and *Research* trigger *Knowledge*, which triggers sending *Order* to *Production* of fire extinguishers.

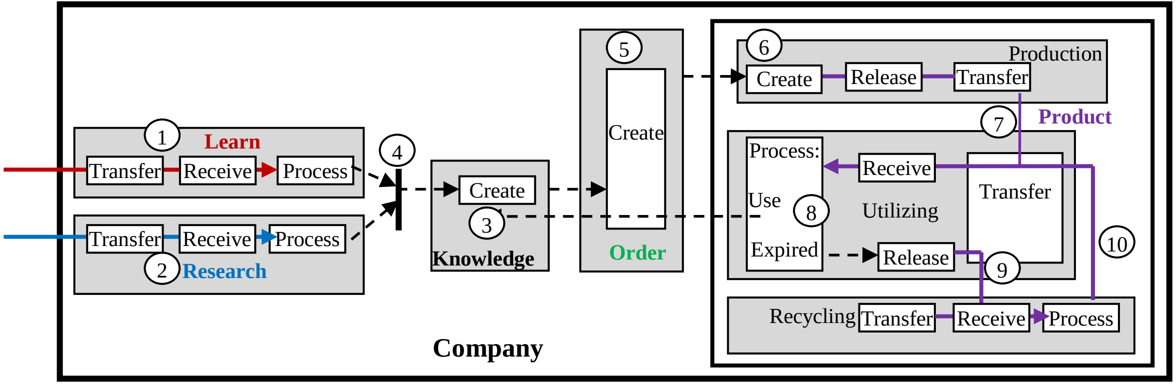


Fig. 9. FM representation of a business process for a company that supplies fire extinguishers

What characteristic of such a notion, of a "human-centered model of processes," qualifies it to take us "into the heady world of memes and memetics"? [24]. Further research can throw light on this question; what we have achieved here is the expression of human-centered models of processes while using the same diagrammatic representation for other senses of memes.

C. Agents' Decisions about Meme Adoption and Transmission

Castelfranchi [25] studied agents' decisions with regard to meme adoption and transmission in terms of observed behaviors and actual communication. According to Castelfranchi [25],

- The so called 'contagion' can be the result of decision processes (to believe or not; to adopt or not), and
- The so called spreading can be the result of another possible decision: to pass or not to pass such a meme to others.

The agent is very far from being a passive 'vehicle' of memes; it can actively decide about receiving them and passing them. [25]

Castelfranchi [25] provides a simple model (Fig. 10) of an agent's decisions, using explicit messages aimed at inducing the agent to believe or to act. The agent's decision process is based either on observation or on communication with a source. The agent has to decide,

- whether or not to adopt such a belief, behavior, or method/tool …
- whether or not to pass this meme on to other agents…
- or to actively try to conceal it…

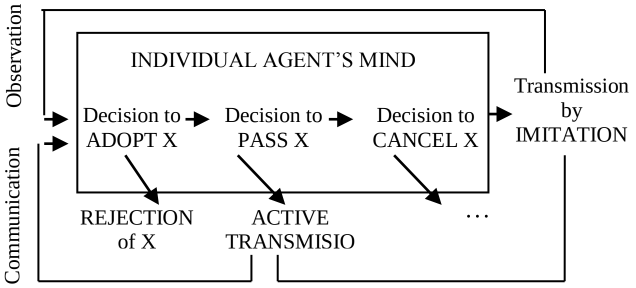


Fig. 10. Partial view of the agent's decision process (redrawn from Castelfranchi [25])

Fig. 11 shows the corresponding FM representation, according to our understanding of the description provided by Castelfranchi [25]. Memes flow from observation (circle 1) and communication (2). Note that the meme sphere in Fig. 11 comprises a hierarchy of three flowsystems, as shown in Fig. 12. Thus, a communication meme "enters" the agent sphere as such, but it can simultaneously be just a meme, the same way positive integers and negative integers can be viewed as just integers.

To emphasize this hierarchy, the meme flowsystem that includes the stages of process (3), transfer (4), receive (5), and de-create (previously called create – to be discussed later) is not drawn in a separate rectangle, as if looking upon it from above (see Fig. 13).

When the meme is processed (3), it triggers (7) the creation of a decision in the flowsystem of decisions as follows:

- "Pass" (8) triggers (9) releasing (10) the communication meme to the outside (11). Note that passing may include multiple instances of copying and passing that are not shown.
- "Adopt" (12) triggers a special type of process that can be called "use" (13).

- “Cancel” (14) triggers *de-creating* (6) the meme. Note that in FM, creation is a type of flow from nonexistence to existence (appearance in the domain of discourse). De-creation is a flow in the opposite direction.

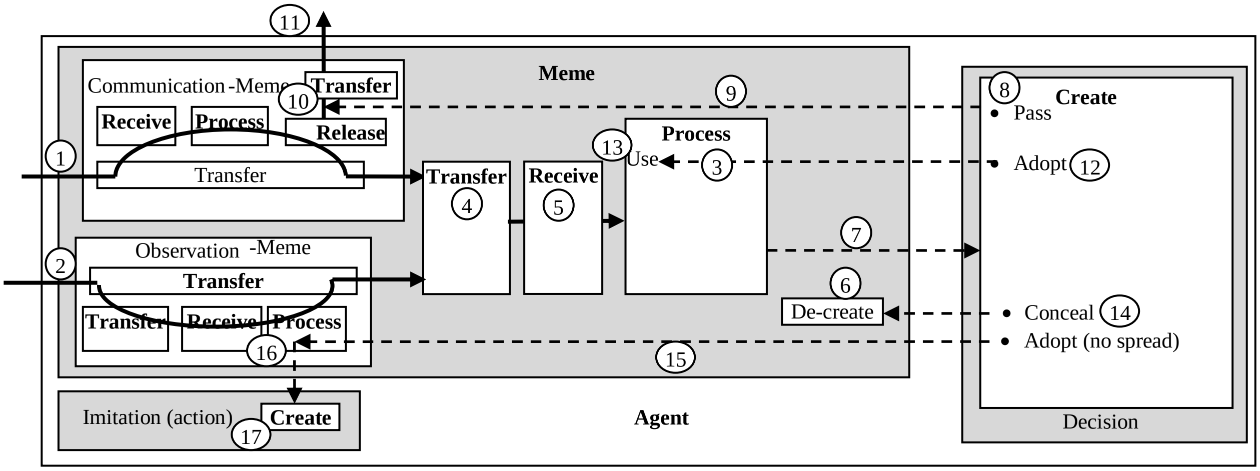


Fig. 11. FM representation of agent’s decision process

- The “Adopt – no spread” of *observation* meme (14) triggers (15) processing of that meme (16), which in turn triggers *acting* the meme (17).

V. Conclusion

This paper has proposed use of the diagrammatic flowthing model (FM) as a uniform language to describe diverse conceptualizations of memes. Such uniformity paves the way in memetics to characterization and categorization of memes and to homogeneity in the field. To demonstrate the viability of this representation method, three examples were selected from different application areas and recast in FM, and the results show the method to be impressive.

The paper also provides a general FM diagram of the flow of memes from a source to infection of a host to (1) flowing to other hosts, (2) triggering some type of reaction, and (3) evolving along the stream of flow.

Future continuing research will represent more types of meme systems to identify common properties that distinguish them from diagrammatic representations of non-meme systems.

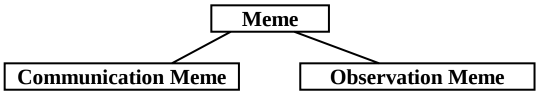


Fig. 12. A meme can be viewed in three subspheres of memes

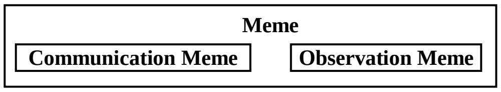


Fig. 13. The three subspheres of memes viewed from above

References

[1] E. Cohen, From ugly duckling to swan: Reconceptualizing information systems as a field of the discipline informing science. Journal of Computing and Information Technology, 7(3), 213-219, 1999.

[2] U. Schmitt, Putting personal knowledge management under the macroscope of informing science. Informing Science Journal, 18, 145-176, 2015.

[3] T. W. Deacon, Memes as signs in the dynamic logic of semiosis: Molecular science meets computation theory, 2003. Retrieved from http://projects.chass.utoronto.ca/semiotics/srb/10-3edit.html

[4] F. Heylighen, Memetics. Principia Cybernetica Web. Nov. 23, 2001. Retrieved from http://pespmc1.vub.ac.be/memes.html

[5] R. Dawkins The selfish gene, new ed. Oxford: Oxford University Press, 1989.

[6] M. B. Prosser, Memetics—a growth industry in US military operations. MS Thesis, School of Advanced Warfighting, United States Marine Corps, 2006. Retrieved from www.dtic.mil/dtic/tr/fulltext/u2/a507172.pdf

[7] J. S. Wilkins, What's in a meme? Reflections from the perspective of the history and philosophy of evolutionary biology. Journal of Memetics: Evolutionary Models of Information Transmission, 2, 1998. http://cfpm.org/jom-emit/1998/vol2/wilkins_js.html

[8] H. C. Speel, A memetic analysis of policy making. Journal of Memetics: Evolutionary Models of Information Transmission, 1, 1997. Retrieved from http://cfpm.org/jom-emit/1997/vol1/speel_h-c.html

[9] I. Shifman, Memes in a digital world: Reconciling with a conceptual troublemaker. Journal of Computer-Mediated Communication, 2013. DOI:10.1111/jcc4.12013

[10] F. Heylighen, What makes a meme successful? In Proceedings, 16th International Congress on Cybernetics (Association Internat. de Cybernétique, Namur), pp. 423-418, 1998.

[11] C. Olah, Memetics [blog posting], 2011. Retrieved from https://christopherolah.files.wordpress.com/2011/01/meme.pdf

[12] H. Bjarneskans, B. Grønnevik, B. and A. Sandberg, A. The lifecycle of memes [web page], 1999. http://www.aleph.se/Trans/Cultural/Memetics/memecycle.html.

[13] G. Morgan, Images of organization. London: Sage, 1997.

[14] Heraclitus In Stanford encyclopedia of philosophy, 2011. Retrieved from http://plato.stanford.edu/entries/heraclitus

[15] D. Hollingsworth, The workflow reference model. Workflow Management Coalition (WFMC), Document No. TC00-1003, Issue 1.1, Jan. 19, 1995. Retrieved from http://www.wfmc.org/Download-document/TC00-1003-The-Workflow-Reference-Model.html

[16] U. M. Husted, A funny thing happened on the way from the forum: The life and death of Internet memes. Ph.D. Thesis, University of Minnesota, 2012.

[17] S. Al-Fedaghi, Schematizing proofs based on flow of truth values in logic. *IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013)*, October 13-16, 2013, Manchester, UK.

[18] S. Al-Fedaghi, S. Flow-based enterprise process modeling. Internatonal Journal of Database Theory and Application, 6(3), 59-70, 2013.

[19] S. Al-Fedaghi, A method for modeling and facilitating understanding of user requirements in software development. Journal of Next Generation Information Technology, 4(3), 30-38, 2013.

[20] S. Al-Fedaghi, Conceptualization of various and conflicting notions of information. Informing Science Journal, 17, 295-308, 2014.

[21] S. Al-Fedaghi, “Flow-based specification of time design requirements,” Int. J. Adv. Comput. Sci. Appl. (IJACSA), vol. 6, no. 8, 2015.

[22] S. Al-Fedaghi, “Diagrammatic representation as a tool in clarifying logical arguments,” Int. J. Adv. Res. Artif. Intell. (IJARAI), vol. 4, no. 10, 2015.

[23] R. Brodie, Virus and the mind: The new science of the meme. Hay House Publishers, 2009. ISBN 978-1-84850-127-0. http://media.evolveconsciousness.org/books/consciousness/Virus-of-the-Mind-The-New-Science-of-the-Meme-Richard-Brodie.pdf

[24] C. Noessel, Whiteboardability: How to make process diagrams memorable. Cooper Journal, May 15, 2008. Retrieved from http://www.cooper.com/journal/2008/05/whiteboardability_portraying_h

[25] C. Castlefranci, Towards a cognitive memetics: Socio-cognitive mechanisms for memes selection and spreading. Journal of Memetics: Evolutionary Models of Information Transmission, 5, 2001. http://cfpm.org/jom-emit/2001/vol5/castelfranchi_c.html

Prediction of Poor Inhabitant Number Using Least Square and Moving Average Method

Ningrum Ekawati
School of Information Management and Computer "AMIKOM Yogyakarta", STMIK AMIKOM Yogyakarta
Yogyakarta, Indonesia

***Abstract*—The number of poor inhabitant in South Kalimantan decreased within the last three years compared with the previous years. The numbers of poor inhabitant differs from time to time. This scaled dynamical number has been a problem for the local government to take proper polices to solve this matter. It will then be necessary to predict a potential number of poor inhabitants in the next year as the basis on subsequent policy making. This research will apply both Least Square and Moving Average method as the measurement to count prediction values. From the results of the study, the prediction analysis by using those two methods is valid for predicting acquired number of poor inhabitant for the next period according to the data from the previous year. Based on the study, the validity of Least Square was 98.35% and Moving Average was 98.79% by using the data in the last seven years.**

Keywords—Poor Inhabitant; Prediction; Least Square; Moving Average

I. Introduction

Poverty is the main problem in South Kalimantan Province, Indonesia [1]. According to Statistic Center Agency, the poor inhabitant is defined as those who averagely spend below the poverty line per capita per month [2]. Based on the Statistic Center Agency in South Kalimantan [2], the number of poor inhabitant decreased each year. In 1999 the number of poor inhabitant was 440,200 and while at the end of 2014 became 182,876 inhabitants. From those data, the number of poor inhabitant decreased 5.28% on average each year and in the last three years decreased 3.31%. There are a decreasing numbers of poor inhabitants in South Kalimantan due to the numbers of building.

The problem was in predicting the number of poor inhabitant in South Kalimantan that decreased in the last three years. However, the prediction cannot be predicted for the next years, and it caused the local government had difficulty to make decision. This study was aimed to decide the way of predicting the number of poor inhabitant in South Kalimantan for the years to come by using Least Square and Moving Average method. Hopefully, the result of the research could help the government to increase the people's life quality.

II. Research Method

Time series analysis was a statistical analysis method applied to predict a future condition. To make an accurate data, the prediction was conducted for a very long time and much data was needed. As one of the choices to describe a future trend, time series analysis can be applied to reflect dynamic variable from one time to another [3]. From the previous studies by using Least Square [4-8] and Moving Average [3][9-14] method, the data and analysis showed the future prediction. It was defined as a management process in making decision. It was described as a prediction process in the unknown future situation. In general term, it was well known as a prediction referring to time series estimation or longitudinal type of data [9].

The Least Square method was often used to predict (Y), due to its detail measurement [4]. The trend line (1) was:

$$Y = a0 + bX \tag{1}$$

$$a = (\Sigma Y) / n$$

$$b = (\Sigma XY) / \Sigma X^2$$

Where:

Y	:	Scaled data (time series) = Trend value prediction.
a0	:	Trend value in the basis year.
b	:	Average growing trend value in each year.
X	:	Time variable (year).

To conduct the calculation, a certain value in time variable (X) was required so that the total variable score was zero or ΣX=0. In analyzing the data with Least Square method, it is generally divided into two parts i.e. "even data" and "odd data" [4].

For odd "n", where:

The interval between two times was one-unit value

It was marked as negative when it was above 0

It was marked as positive when it was below 0

For even "n", where:

The interval between two times gains two-unit value

It was marked as negative when it was above 0

It was marked as positive when it was below 0

Generally, linier line equation from time series analysis (2) was:

$$Y = a + bX \tag{2}$$

Description:

Y is a variable that trend was searched.

X is a time variable (year).

Meanwhile, to find constant value "a" and parameter value "b" (3) was:

$$a = \Sigma Y / n, \text{ and } b = \Sigma XY / \Sigma X^2 \quad (3)$$

Moving Average method was a prediction approach by taking some observed groups of the values, finding the average, and using the average values as a prediction of subsequent period. The formula (4) was [10]:

$$F_t = \frac{A_{t-1}+A_{t-2}+A_{t-3}+A_{t-n}}{n} \quad (4)$$

Description:

F_t : Forecast for the coming period
n : Number of period to be averaged "n"

$A_{t-1} + A_{t-2} + A_{t-3}$: Actual Occurrences in the past period, two period ago, three period ago, and so on respectively.

III. Results and Analysis

Based on the Statistic Central Agency, the number of poor inhabitant in South Kalimantan from 1996 and not for every year due to the calculation until the Province level for once in three year. So, the data in 1996, 1999, and 2000 was counted every year [2]. So, the prediction calculation of poor inhabitant in this study focused on time year variable. Time series analysis with Least Square and Moving Average can be applied to identify time year variable.

The required data of the research were those of poor inhabitant in South Kalimantan Province starting from 1999 to 2014 [2].

TABLE I. The Data of Poor Inhabitant from 1999 to 2014

No	Year	Total
1	1999	440,200
2	2000	385,300
3	2001	357,500
4	2002	259,800
5	2003	259,000
6	2004	231,000
7	2005	235,700
8	2006	278,451
9	2007	233,500
10	2008	218,898
11	2009	175,977
12	2010	181,963
13	2011	194,623
14	2012	190,597
15	2013	184,297
16	2014	182,876

A. *Least Square Method*

In this study, the data for the "odd data", previous collected data from the last nine years were required. Meanwhile, when processing data tabulation for an "even data", the previous data collection from the last ten years are required.

1) "Odd Data": Before measuring the prediction of poor inhabitant in 2015, the test was conducted to the number of poor inhabitant in 2008, 2009, 2010, 2011, 2012, 2013, and 2014 to know whether Least Square was valid or not. Compared with the data of poor inhabitant in the last seven years. To find the data of poor inhabitant in 2008, the data of poor inhabitant in 1999 to 2007 was used. The next was finding the value of X, XY, and X^2.

TABLE II. Variable Data of Poor Inhabitant from 1999 to 2007

No	Year	Total (Y)	X	XY	X^2
1	1999	440,200	-4	-1,760,800	16
2	2000	385,300	-3	-1,155,900	9
3	2001	357,500	-2	-715,000	4
4	2002	259,800	-1	-259,800	1
5	2003	259,000	0	0	0
6	2004	231,000	1	231,000	1
7	2005	235,700	2	471,400	4
8	2006	278,451	3	835,353	9
9	2007	233,500	4	934,000	16
Total		2,680,411		-1,419,747	60

Thus, to find "a" value was:

$a = \Sigma Y / n$

$a = 297{,}823.44$

And to measure "b" value was:

$b = \Sigma XY / \Sigma X^2$

$b = -23{,}662.45$

After gaining the "a" and "b" value, the equation line was:

$Y = a + bX$ (in 2008 the X score was 5)

After measuring the linear line, the number of poor inhabitant in 2008 was:

$Y = 179{,}511.192$

It means that the total number of poor inhabitant in 2008 was 179,511 inhabitants.

The next phase was finding the total number of poor inhabitant in 2009. The data of poor inhabitant in 2000 to 2008 was collected. With the same calculation, in 2009 the result was 187,937 inhabitants. The data of poor inhabitant in 2001 to 2009 was used to identify the number of poor inhabitant for 2010. With the same calculation, the number of poor inhabitant in 2010 was 178,954 inhabitants. The data in 2002 to 2010 was used to know the number of poor inhabitant in 2011. Also, with the same calculation, the total number of poor inhabitant in 2011 was 181,580 inhabitants. The data in 2003 to 2011 was used to identify the number of poor inhabitant in 2012. With the same calculation, the number of poor inhabitant in 2012 was 174,609 inhabitants. The data in 2004 to 2012 was applied to find the number of poor inhabitant in 2013. The same calculation showed that the number of poor inhabitant in 2013 was 171,023 inhabitants. The data in 2005 to 2013 was applied to identify the number of poor inhabitant in 2014. With the same calculation, the total number of poor inhabitant in 2014 was 161,790 inhabitants.

If the different score between Least Square method was >40%, it was considered to be invalid. Compared to the accurate score in 2008, the different was 17.99% (39,387 inhabitants) and the data was valid. In 2009, the different was 6.80% (11,960 inhabitants), it means that the data was valid as well. In 2010, the different was 1.65% (3,009 inhabitants), it means that the data was valid. In 2011 the difference was 6.70% (13,043 inhabitants), it is also means that the data was valid. In 2012, the comparison was 8.39% (15,988 inhabitants), the data also was valid. In 2013, the difference was 7.20%

(13,274 inhabitants), the data was considered to be valid. And in 2014, the comparison was 11.53% (21,086 inhabitants) it means that the data was valid as well. Based on the seven differences, the all data was valid. So, Least Square method was effective or accurate.

The next phase was calculating the prediction number of poor inhabitant in 2015. Based on the data tabulation for "odd data", the poor inhabitant data were needed from the last nine years, starting from 2006 to 2014.

TABLE III. VARIABLE DATA OF POOR INHABITANT FROM 2006 TO 2014

No	Year	Total (Y)	X	XY	X^2
1	2006	278,451	-4	-1,113,804	16
2	2007	233,500	-3	-700,500	9
3	2008	218,898	-2	-437,796	4
4	2009	175,977	-1	-175,977	1
5	2010	181,963	0	0	0
6	2011	194,623	1	194,623	1
7	2012	190,597	2	381,194	4
8	2013	184,297	3	552,891	9
9	2014	182,876	4	731,504	16
Total		1,841,182		-567,865	60

The table 3, showed that the "a" and "b" values were obtained. To count "a" and "b" values, the following formula was applied:

To find out "a" value was:

$a = \Sigma Y / n$

$a = 204{,}575.8$

And to find "b" value was:

$b = \Sigma XY / \Sigma X^2$

$b = -9{,}464.42$

After "a" and "b" values were obtained, the linear line was found as follows:

$Y = a + bX$ (for year 2015 the value of X is 5)

After finding the linear measurement, the number of poor inhabitant in 2015 was as follows:

$Y = 157{,}253.7$

It means that the number of poor inhabitant in 2015 was 157,254 inhabitants.

So, the trend analyzing graphic with Least Square method for the different result and the prediction result in the last seven years were:

TABLE IV. THE DATA OF POOR INHABITANT BY LEAST SQUARE METHOD

Year	Actual	Prediction
2008	218,898	179,511
2009	175,977	187,937
2010	181,963	178,954
2011	194,623	181,580
2012	190,597	174,609
2013	184,297	171,023
2014	182,876	161,790
2015		157,254

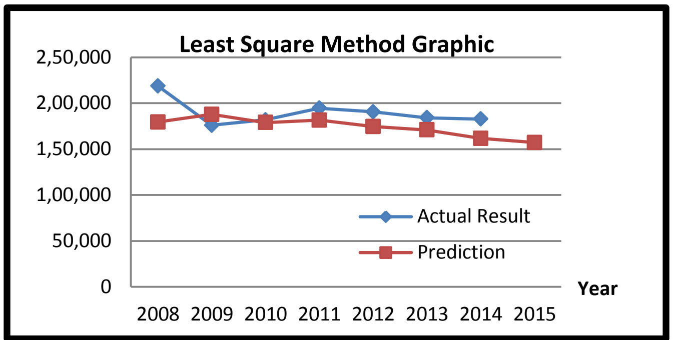


Fig. 1. Graphic Prediction by Using Least Square Method

2) "Even Data": the required data in the case of "even data" are those of poor inhabitant in South Kalimantan starting from 2005 to 2014.

TABLE V. VARIABLE DATA OF POOR INHABITANT FROM 2005 TO 2014

No	Year	Total (Y)	X	XY	X^2
1	2005	235,700	-9	-2,121,300	81
2	2006	278,451	-7	-1,949,157	49
3	2007	233,500	-5	-1,167,500	25
4	2008	218,898	-3	-656,694	9
5	2009	175,977	-1	-175,977	1
6	2010	181,963	1	181,963	1
7	2011	194,623	3	583,869	9
8	2012	190,597	5	952,985	25
9	2013	184,297	7	1,290,079	49
10	2014	182,876	9	1,645,884	81
Total		2,076,882		-1,415,848	330

Based on table 5, "a" and "b" values were obtained. To find those scores, the following formula was applied:

To find the "a" value was:

$a = \Sigma Y / n$

$a = 207{,}688.2$

To find the "b" value was:

$b = \Sigma XY / \Sigma X^2$

$b = -4{,}290.45$

After the values of "a" and "b" was gained, the linear measurement was as follows:

$Y = a + bX$ (in 2015 the value of X was 11)

With that equation, the number of poor inhabitant in 2015 was as follows:

$Y = 160{,}493.3$

It means that the prediction number of poor inhabitant was 160,493 inhabitants.

From the calculation of Least Square method, the prediction number of poor inhabitant in 2015 for "odd data" was 157,254 inhabitants. And for the "even data" was 160,493 inhabitants. So, the different was 2.02%. However, the result of prediction number of poor inhabitant could be wrong due to some cases such as natural disaster, disease epidemic, etc.

B. Moving Average Method

Before measuring the prediction number of poor inhabitant in 2015, the test of number of poor inhabitant in 2008, 2009, 2010, 2011, 2012, 2013, and 2014 by using Single Moving Average method were conducted to know whether the data was valid or not compared with the accurate data of poor inhabitant in the last seven years. To count the prediction number in 2015, the prediction number in 2014 was counted at first.

TABLE VI. THE DATA OF POOR INHABITANT FROM 1999 TO 2014

No	Year	Total
1	1999	440,200
2	2000	385,300
3	2001	357,500
4	2002	259,800
5	2003	259,000
6	2004	231,000
7	2005	235,700
8	2006	278,451
9	2007	233,500
10	2008	218,898
11	2009	175,977
12	2010	181,963
13	2011	194,623
14	2012	190,597
15	2013	184,297
16	2014	182,876

The next phase was defining the value of poor inhabitant by using Single Moving Average method for two periods as follows:

TABLE VII. VARIABLE DATA OF POOR INHABITANT BY SINGLE MOVING AVERAGE TWO PERIOD

No	Year	Actual	Prediction
1	1999	440,200	-
2	2000	385,300	-
3	2001	357,500	412,750
4	2002	259,800	371,400
5	2003	258,960	308,650
6	2004	231,000	259,400
7	2005	235,700	245,000
8	2006	278,451	233,350
9	2007	233,500	257,076
10	2008	218,898	255,976
11	2009	175,977	226,199
12	2010	181,963	197,438
13	2011	194,623	178,970
14	2012	190,597	188,293
15	2013	184,297	192,610
16	2014	182,876	187,447
17	2015	-	183,587

If the different between prediction calculation with Single Moving Average for two periods with the results was >40%, then it is considered to be invalid.

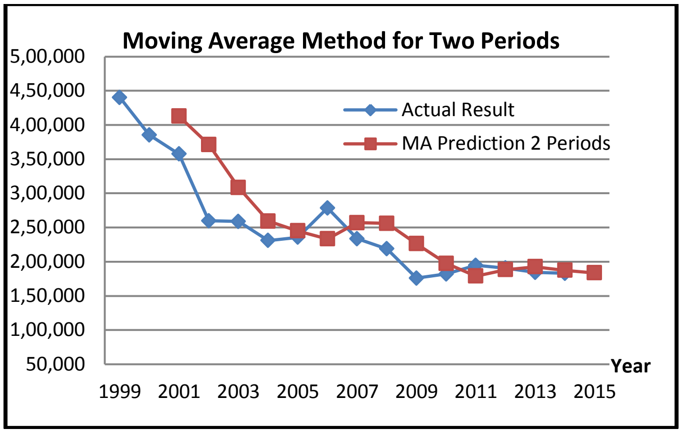


Fig. 2. Graphic Prediction by Using Moving Average Method for Two Periods

Compared to the accurate data in 2008, the different was 14.48% (37,078 inhabitants), it was considered to be valid. In 2009, the different was 22.20% (50,222 inhabitants), it means that the result was considered to be valid. In 2010, the different was 7.84% (15,475 inhabitants), it means that the data was valid. In 2011, the comparison was 8.04% (15,653 inhabitants), the data was valid. In 2012, the difference was 1.21% (2,304 inhabitants), the data was also valid. While in 2013, the difference was 4.32% (8,313 inhabitants) and the data was valid. Lastly, in 2014 the difference was 2.44% (4,571 inhabitants) it also means the data was valid. According to those seven comparisons, the use of Single Moving Average was effective.

C. The Comparison Result of Least Square and Moving Average Method

According to Least Square and Moving Average method, if the difference between prediction calculation with the result was >40% it means that the data were invalid. Based on table 8, there was a comparison result between Least Square and Moving Average method for the last seven years.

TABLE VIII. THE COMPARISON DATA OF LEAST SQUARE METHOD AND SINGLE MOVING AVERAGE METHOD FOR TWO PERIODS

Year	Actual	Prediction		Difference (%)		Result
		Least Square	*Moving Average*	*Least Square*	*Moving Average*	
2008	218,898	179,511	255,976	17.99	14.48	Valid
2009	175,977	187,937	226,199	6.80	22.20	Valid
2010	181,963	178,954	197,438	1.65	7.84	Valid
2011	194,623	181,580	178,970	6.70	8.04	Valid
2012	190,597	174,609	188,293	8.39	1.21	Valid
2013	184,297	171,023	192,610	7.20	4.32	Valid
2014	182,876	161,790	187,447	11.53	2.44	Valid
2015	-	157,254	183,587			

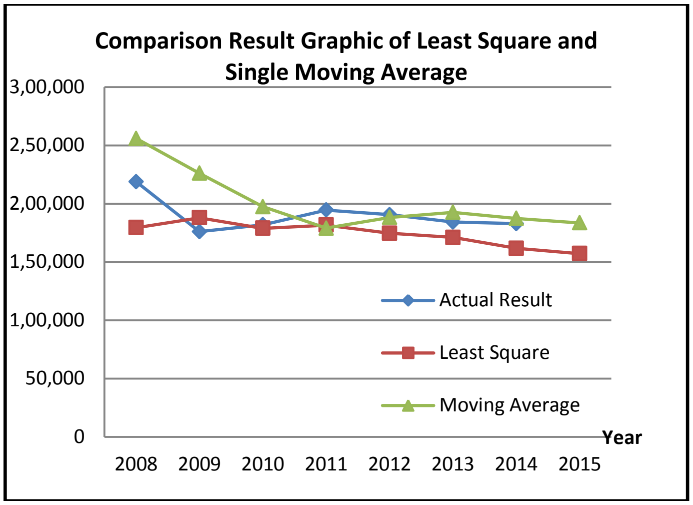


Fig. 3. Comparison Result Graphic of Least Square and Single Moving Average Method for Two Periods

In accordance with the figure 3, the lower data for Least Square method was 1.65% and 1.21% for Moving Average method, so the data was considered to be effective. The validity of Least Square and Moving Average was based on the accurate measured data.

IV. Conclusion

The use of Least Square and Single Moving Average method was effective to predict the number of poor inhabitant in South Kalimantan for the next period.

From the result of the prediction calculation, the number of poor inhabitant in 2008 was 179,511 inhabitants. And in 2009 was 187,937 inhabitants. While in 2010 was 178,954 inhabitants. In 2011 was 181,580 inhabitants, in 2012 was 174,609 inhabitants, in 2013 was 171,023 inhabitants, and in 2014 was 161,790 inhabitants. If the difference between the calculation of Least Square and the accurate result was >40%, it was considered to be invalid. Compared to the actual result in 2008, the difference was 17.99% (39,387 inhabitants), it means that the result was valid. In 2009, the difference was 6.80% (11,960 inhabitants), it means that the result was valid. In 2010, the comparison was 1.65% (3,009 inhabitants), it also considered to be valid. In 2011, the comparison was 6.70% (13,043 inhabitants) and the data was valid. In 2012, the difference was 8.39% (15,988 inhabitants) and the data was valid. In 2013, the comparison was 7.20% (13,274 inhabitants) it means that the data was valid. And in 2014, the difference was 11.53% (21,086 inhabitants) and the data was considered to be valid. It means that Least Square method was approximately effective.

From the calculation result from prediction of poor inhabitant in 2008 was 255,976 inhabitants, in 2009 was 226,199 inhabitants, in 2010 was 197,438 inhabitants, in 2011 was 178,970 inhabitants, in 2012 was 188,293 inhabitants, in 2013 was 192,610 inhabitants, in 2014 was 187,447 inhabitants. It was considered to be invalid when the difference between the calculation with Single Moving Average for two periods and the accurate result was >40%. Compared to the accurate data in 2008 the difference was 14.48% (37,078 inhabitants), it means that the result was valid. In 2009, the difference was 22.20% (50,222 inhabitants), and the data was valid. In 2010, the difference was 7.84% (15,475 inhabitants), it means that the result was also valid. In 2011, the comparison was 8.04% (15,653 inhabitants) it means that the data was valid. In 2012 the difference was 1.21% (2,304 inhabitants), and the data was considered to be valid. In 2013, the comparison was 4.32% (8,313 inhabitants), it also considered to be accurate. In 2014, the difference was 2.44% (4,571 inhabitants) it means that the data was valid. Based on the seven comparisons, the all data was accurate or valid. Thus, Single Moving Average was approximately effective.

The accurate result of Least Square was 98.35% and 98.79% for Moving Average, so it was considered to be valid in predicting the number of poor inhabitants.

For the next researches, the number of data and additional variable are required. Smart system can be used as a method to predict the number of poor inhabitant.

References

[1] F. Amina and M. I. Irawan, “Prediksi Jumlah Penduduk Miskin di Kalimantan Selatan Menggunakan Jaringan Syaraf Tiruan Backpropagation,” 2014.

[2] Statistic Center Agency of South Kalimantan Province: http://kalsel.bps.go.id/Subjek/view/id/23#subjekViewTab1|accordion-daftar-subjek1

[3] I. Majerová and T. Pražák, “Estimation of Economic Development in Papua New Guinea: Linear Trend Analysis or Moving Average Model?,” *Procedia - Soc. Behav. Sci.*, vol. 110, pp. 450–460, 2014.

[4] M. I. F. Rambe, “Perancangan Aplikasi Peramalan Persediaan Obat-Obatan Menggunakan Metode Least Square (Studi Kasus: Apotik Mutiara Hati),” pp. 49–53, 2014.

[5] H. Meilin and X. Yanxia, “Estimation of the complex frequency of a harmonic signal based on a linear least squares method,” *Geod. Geodyn.*, vol. 6, no. 3, pp. 220–225, 2015.

[6] S. A. Korkmaz and M. Poyraz, “Least Square Support Vector Machine and Minumum Redundacy Maximum Relavance for Diagnosis of Breast Cancer from Breast Microscopic Images,” *Procedia - Soc. Behav. Sci.*, vol. 174, pp. 4026–4031, 2015.

[7] S. Wang and W. Shang, “Forecasting direction of China security index 300 movement with least squares support vector machine,” *Procedia Comput. Sci.*, vol. 31, pp. 869–874, 2014.

[8] G. Tan, J. Yan, C. Gao, and S. Yang, “Prediction of water quality time series data based on least squares support vector machine,” *Procedia Eng.*, vol. 31, pp. 1194–1199, 2012.

[9] L. Abdullah, “ARIMA Model for Gold Bullion Coin Selling Prices Forecasting,” vol. 1, no. 4, 2012

[10] R. Kumar and D. Mahto, “Application of Proper Forecasting Technique in Juice Production: A Case Study,” vol. 13, no. 4, 2013.

[11] V. Ruiz, M. A. Pérez, and A. Olasolo, “Dynamic Portfolio Management Strategies based on the Use of Moving Averages,” Procedia - Soc. Behav. Sci., vol. 109, pp. 1277–1281, 2014.

[12] K. Mivule and C. Turner, “Applying Moving Average Filtering for Non-interactive Differential Privacy Settings,” Procedia Comput. Sci., vol. 36, pp. 409–415, 2014.

[13] H. K. Yu, N. Y. Kim, S. S. Kim, C. Chu, and M. K. Kee, “Forecasting the Number of Human Immunodeficiency Virus Infections in the Korean Population Using the Autoregressive Integrated Moving Average Model,” Osong Public Heal. Res. Perspect., vol. 4, no. 6, pp. 358–362, 2013.

[14] A. Pal, J. P. Singh, and P. Dutta, “The Path Length Prediction of MANET Using Moving Average Model,” Procedia Technol., vol. 10, pp. 882–889, 2013.

Study of Automatic Extraction, Classification, and Ranking of Product Aspects Based on Sentiment Analysis of Reviews

Muhammad Rafi
Computer Science Department
National University of Computer and Emerging Sciences
Karachi, Pakistan

Usama Noman
Computer Science Department
National University of Computer and Emerging Sciences
Karachi, Pakistan

Muhammad Rafay Farooq
Computer Science Department
National University of Computer and Emerging Sciences
Karachi, Pakistan

Abdul Rehman Farooq
Computer Science Department
National University of Computer and Emerging Sciences
Karachi, Pakistan

Umair Ali Khatri
Computer Science Department
National University of Computer and Emerging Sciences
Karachi, Pakistan

***Abstract*—It is very common for a customer to read reviews about the product before making a final decision to buy it. Customers are always eager to get the best and the most objective information about the product theywish to purchase and reviews are the major source to obtain this information. Although reviews are easily accessible from the web, but since most of them carry ambiguous opinion and different structure, it is often very difficult for a customer to filter the information he actually needs. This paper suggests a framework, which provides a single user interface solution to this problem based on sentiment analysis of reviews. First, it extracts all the reviews from different websites carrying varying structure, and gathers information about relevant aspects of that product. Next, it does sentiment analysis around those aspects and gives them sentiment scores. Finally, it ranks all extracted aspects and clusters them into positive and negative class. The final output is a graphical visualization of all positive and negative aspects, which provide the customer easy, comparable, and visual information about the important aspects of the product. The experimental results on five different products carrying 5000 reviewsshow 78% accuracy. Moreover, the paper also explained the effect of Negation, Valence Shifter, and Diminisher with sentiment lexiconon sentiment analysis, andconcluded that they all are independent of the case problem, and have no effect on the accuracy of sentiment analysis.**

Keywords—Aspect ranking; Product Aspect Ranking; Sentiment analysis; Sentiment lexicon

I. INTRODUCTION

Web 2.0 is rich on user-generated contents for different products. For example, CNet.com involves more than seven million product reviews, whereas Pricegrabber.com contains millions of reviews on more than 32 million products. Consumers nowadays believe to buy more products from online store than any physical score. Ever since the number of customers who prefer to shop online increases, the phenomenon of askingreviews about the product from family and friend also increased. Now with the advancement of technology, the task of asking for reviews from friends and family has shifted to Product Reviews website. Seventy percent consumers believe that the online reviews are the most trusted and reliable source of information [1]. Although web has a large collection of information, sifting through these texts and extracting valuable information from these disorganized reviews is very challenging and daunting task, but can be solved using sentiment analysis.

Sentiment analysis carries great importance in digital world. Identifying sentiments from the natural text is not very difficult but tricky. Correct analysis of reviews could increase the sale of a company to 200% in a month; therefore, many researchers are experimenting with different methods to find the complete solution of this daunting task. The main challenge in finding sentiments from the text is to find its scope and its intensity. Since mostly reviews are subjective and carry ambiguous opinion, it is very hard to find opinion words, understand their contextual meaning, and identify their scope. Due to non-uniform structure of web, it is also very challenging to mine the information the reader actually needs. Online reviews constitute a small part of this huge clustered of web pages.

II. LITERATURE REVIEW

The rapid growth in e-commerce is due to increasing trust of online customers. There are millions of products, from thousands of manufactures and distributors available for sale online. Every category has hundreds of product to choose from, and it is very difficult for an online buyer to make a wisedecision. Therefore,buyers go through reviews about the product to make a final decision, but due to subjectivity and ambiguity of reviews, it often does not reach any valuable conclusion. The difficult part of this activity is to search this distributed information from multiple website;analyze the subjectivity of text, and conclude the final notions about the product from these reviews.

Recent study [2]showed the impact of reviews on the sale of product, and opinion of the customer. In this work, a special summarization technique is presented which was different from traditional text summarization because it focuses only on important aspects of the product rather than summarization of the whole review. This summary proved to be very beneficial for the online users who are about to make a purchase. Another recent work is of Zheng-Jun Zha,Jianxing Yu, Meng Wang, and Tat-Seng Chua [3]related to the identification of product aspects. The work proposed a framework that can rank the important aspects of the product by exploiting product's aspects frequency and its probability in a review. Their algorithm showed significant improvements over other already proposed methods when tested on 95 thousands customer reviews. Identifying the sentiments associated with different aspect of product seems to be a very instrumental in overall sentiment of the product.

Some researchers developed their own sentiment lexicons, and devised tools that could automatically extract reviews, and find important features based on supervised learning techniques. Bo Pang, Lillian Lee, and Shivakumar were among the very few early researchers who proposed the technique of sentiment classification using machine learning [4]. They proposed a method, which can find the sentiment of a document not by topic but by overall sentiment i.e. if the review is positive or negative. They also highlighted the importance of unigram words in sentiment classification technique in their paper. Hanhoon Kang, SeongJoonYoo,and Dongil Han proposed a new lexicon for sentiment classification because of lack of sentiment words in already existing sentiment corpuses [5]. Their focus was to narrow the classification accuracy gap between positive and negative sentiment documents. They proposed a modified Naïve Bayes algorithm, which narrowed down the classification gap to 3.6% as compared with original Naïve Bayes. They chose the dataset of restaurant reviews for their experiment and concluded that unigrams and bi-grams features play a major role in sentiment analysis of reviews. Likewise, Kushal Dave, Steve Lawrence, and David M. Pennockdeveloped an opinion-mining tool that can distinguish between positive and negative reviews automatically by assigning the features some scores based on heuristics [6]. The tool was prune to web based searches due to noise and ambiguity, and used supervised learning to find the important aspects from the text. In addition to that, Michael Wieg et.al presented a concrete summary in his paper on the role of negation in Sentiment Analysis [7]. They present computational approaches, and modeled the role of negation in sentiment analysis. In addition, they also discussed the limitation and challenges in negation modeling followed by the detection and scope of negative words. In comparison, some researchers discussed the impact of irony and sarcasm in online reviews and their effect on sentiment analysis task [8]. Few of them also explained the influence of negative and valence shifter words in determining the overall sentiment of a review. Elena Filatova presented a corpus generation experiment that can identify irony and sarcasm from a corpus in two levels: document level and text utterance where text utterance can range from a single sentence to a complete document [9]. Livia Polanyi and Annie Zaenenworked on valence shifters to determine the attitude of writers towards the material being described [10]. In contrast, Alistair Kennedy and Diana Inkpen use the help of contextual shifter to classify movie reviews [11]. They specifically examine three types of contextual valence shifters namely negations, intensifier, and diminisher, and studies their effect on classification of reviews. They did not assign weights to negative and positive words and treated all the words on the same level.

Our research mainly focused on how to identify important aspects of a product from its reviews, and rank those aspects based on their sentiment scores. Our work closely relates to the work of Zheng-Jun Zha [12], but instead of working on term frequency and probability, we used different lexicons to identify the score of aspect and later rank those aspects based on their sentiment scores. In this paper, we are discussing a lexicon-based approach to find the sentiment score of aspects. The impact of sentiment lexicon on negation handling which is the most important part of sentiment analysis is also discussed. In summary, the main contribution of this research is as follows:

i) We proposed a lexicon based approach to find the sentiment of product (aspect).

ii) We analyzed the importance of a good opinion lexicon, and its effect on negation-handling task. Wealso concluded that if a good opinion lexicon is used, then we do not need to handle important features of linguistics like Valence shifter in the task of sentiment analysis.

Next section describes the proposed approach of finding the sentiment of aspects through lexicon. It describes the pre-processing task, challenges of extracting text from reviews, and the task of identifying relevant aspects and its impactonresults. After that, the paper has the intermediate results of different lexicon with different window sizes. The final section describes negation handling, the effect of Valence shifter and Diminisher on sentiment scores and its effect on lexicon based approach.

III. The Proposed Approach

In this section, we explained our proposed approach and working of the system. Before proposing our solution, we also presented a short summary of relevant challenges and problems faced previously. Our system has four phases namely Extraction of Reviews from Web, Identifying Aspects of Product from those reviews, Sentiment Analysis of Product (Aspect), and finally the ranking of product aspects based on sentiment score. In each phase, the output of one phase feed acts as an input for another. First, we extract reviews from the web and identify aspects by detecting noun phrases [13]. After finding aspects, we do sentiment analysis of wordssurrounded by those aspects, and assign score to each aspect. Finally, we rank all aspects based on their sentiment scores and present this information to the user. The following section will explain each phase of the system in detail.

A. *Pre-Processing of Text*

Preprocessing of text plays a vital role in the area of text classification and natural language processing. In order to get good results, this step plays a very important role in our system. The impact of pre-processing in the field of text classification is extensively studied, and research on various languages like Arabic, Turkish, and Portuguese [14], [15], [16] support our motivation behind doing pre-processing at this step. It has already proven that preprocessing takes almost 80% of the total time in classification process [17]. Many good techniques like TF/IDF, Stop word removal and stemming showed considerable impact on classification accuracy of documents with different domain dataset [18]. Experiments also conclude that different combination of preprocessing techniques should be applied instead of enabling or disabling them all to increase the accuracy [19]. In our approach, we removed stop words that expand sentient word's domain and enhance discrimination degree between documents.

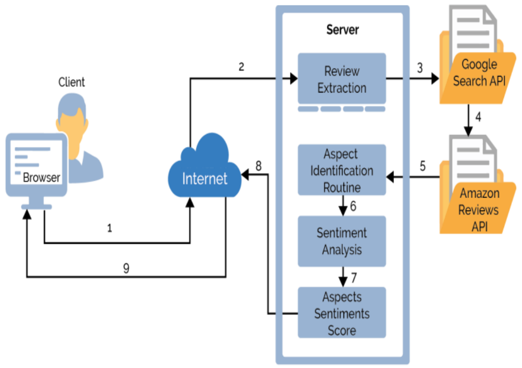


Fig. 1. Proposed framework of the system

B. *Extraction of Reviews from Web*

The diversity of Web 2.0makes the process of extracting relevant information from this unstructured and non-uniform spider of pages a very daunting task. For many natural language processing tasks, the size and quality of data used for training and testing is very crucial. Due to rapid increase of data around the web, it is very important to study how only the relevant data from different web sources can be extracted and processed for relevant tasks. Since most web pages contain tags and other non-content HTML characters [20], it is easy to extract content from any web page if we can exploit these tags. In our approach, we exploited HTML and XTML tags and used python library for web scrapping to extract the reviews from different websites.

C. *Aspect Identification*

Identifying important and relevant aspects of the product is very important and the most sensitive part in sentiment analysis process. Proper identification of aspects is very challenging due to diversity of Natural Language Text. The most novel approach to identify important aspects from online reviews is to observe customer reviews and sort the frequency of aspects mentioned in most of the reviews. The majority class will represent the important aspect of that product [12]. For aspect identification, we find all the frequent nouns from the text [21] and sort them with their term frequency. For unigram and bi-gram aspects, we filter all Nouns with their term frequency. In addition, if we have two aspects with one having a common sub-string of other, we discard the shorter length aspect to be more specific. For example if we have two aspects "battery" and "battery life," we will chose the second one because it is more specific than first. Table below shows some extracted unigram and bigram aspects of **"IPod"** by our system.

TABLE I. UNIGRAMS OF IPOD

Battery	Songs	Music	Capacity	Software
Device	I-Tunes	Service	Computer	I-Pod
Controls	Interface	Audio	Bass	Adapter
Click Wheel	Sound Quality	Battery life	Storage capacity	Click Wheel
Battery replacement	Sleek design	Better technology	Color display	Battery replacement

D. *Sentiment Analysis*

Sentiment analysis and opinion mining are the most widely researched topics in the domain of Natural Language and Data mining. Different researchers have worked on different layers of sentiment analysis from Document level to sentence level and aspect level [22]. The important challenge faced in the area of sentiment analysis is the true meaning of opinion words. Likewise, it is not always necessary that an opinion word like *"good"* always carries some positive opinion when appears in a sentence. For example, the sentence *"I want to buy a good camera, can you please give me some suggestion"* does not carry any opinion about any camera but has opinion words. In our proposed methodology of finding the sentiment score of identified aspects, we used lexicon-based approach. We have experimented with different lexicons and tested different combinations to maximize the classification accuracy of sentiments. We used lexicons of **NRC Canada**[23]also known as Senti140**,** and **CSUIC** also known as Opinion Lexicon by Minqing Hu et al.[13] in order to find the sentiment scores of the words surrounded by the aspect**.** The following section will explain the intermediate results and analysis when different lexicons are experimented on and tested.

Data Set

In our experiment, we have used the dataset of "Ming Lui spam detection in fake reviews" [24]. The dataset contains nine products with more than 5000 reviews in the whole collection. The reviews are manually annotated with ranking of each aspect identified. We train our algorithm in five products reviews, and find generalization accuracy. The format of the review is ASPECT [[+/-] RANK] ##REVIEW. Reviews were annotated and with each review, its aspect score was given. In addition to that, we also experiment with SemEval 2014 data set of Restaurant and Laptop reviews[25].

The dataset of each category is divided into two partsnamely Training and Trialing. Laptop Training dataset contains 1900 reviews, whereas restaurant training dataset contains 1350 reviews all annotated by experienced annotators.

IV. Experiments

In our approach, we used distance based linking to find how the opinion words affect the meaning of the target aspect. A windows size of four and opinion words from the target aspect on both sides are exhausted. We have applied distance based linking to aspect level and with each aspect; we stored the surrounding opinion phrase with window size of four. If there are more opinion words than aspects, then we apply the score of those opinion words to all the aspects coming in the window. For example, if the review is *"The camera is very good and gives amazing and astonishing results."* There are three opinion words *very good, amazing, astonishing,* but only one aspect *camera.* Since all these opinion words come on the window size of four with camera, they will modify the sentiment score of camera. Following section will explain all the intermediate results of our experiment with different lexicons and window sizes.

First, we used only AFINN as sentiment lexicon and calculated sentiment scores of surrounding words in thedistance of two. Our algorithm gave average accuracy of 26%. The major reason of this low accuracy was the short length of this lexicon. There were only 2800 words in AFINN and without handling negation and bi-words, the accuracy become worst. As a result, we add another rich lexicon named CSUIC aka Opinion Lexicon compiled by Bing Lui [24]. The results showed 6% improvement over AFFIN with overall accuracy to 32%. It is important to note that the window size of experiment was set to two during this analysis. After experimenting with windows size of two, we gradually increased our windows size from two to three and then to four. Major improvements are observed when window size was set to four.Mostly reviews contain long sentences due to which the window size of five decreased the accuracy. Moving further, we also experimented with different other lexicons like Sentiment 140 apart from CSUIC and AFFIN, and finally found the three best lexicons.

TABLE II. Affin, Csuic, and Senti140 Score with Window Size = 3

Products	Distance 3	
	AFFIN+CSUIC+Sennti140	CSUIC+Senti140
IPod	42.7%	45.8%
CanonG3	32.9%	34.3%
Hitachi Router	44.2%	47.2%
Norton	29.5%	30.3%
Average	37.7%	40.4%

TABLE III. Affin, Csuic, and Senti140 Score with Window Size = 4

Products	Distance 4 (Unigram)		Bigram
	AFFIN+CSUIC+Senti140	CSUIC+Senti140	
IPod	53.1	53.6	68.75
CanonG3	36.0	36.0	52.79
Hitachi Router	52.1	52.5	63.77
Norton	31.6	32.8	54.91
Average	49.9	49.4	60

After experimenting with different lexicons with varying window sizes, the results concluded the importance of a good lexicon for sentiment analysis. We further observed that having a rich lexicon like Senti140, the addition of AFFIN was useless as the scores of AFFIN were replicated in Senti140. Therefore, we removed AFFIN from our customized lexical dictionary. When we exploit unigram features, we achieved 49% accuracy with window size = 4

A. Negation

Identifying and handling negation is the most important step in sentiment analysis task. The most challenging part of handling negation is to identify its scope and its effect on sentiment words. In addition, it is not yet cleared that how the effect and resolution of negation should be represented and generalized for different domains [26]. Different experiments showed that the identification and handling of negation improves both accuracy and performance of sentiment analysis system [27]. In our proposed approach, we handled Unigrams and Bigrams with different windows sizes using different lexicons. We have achieved the highest accuracy of 60% with lexicon CSUIC and Senti140 with Window size of four. To handle negation, we used customized list, which contains all the negative words. Previously, for all negative words we were using the score of Senti140 but after writing a separate routine for negative words, our accuracy shoots up to 10% from previous results. If any opinion word matched with the negative list word, we appended the word NOT to all the surrounding opinion words in the array with the window size of four. For example, *"I do not like IPhone5"* was converted into *"I do not NOT_like NOT_IPhone5."* The average accuracy of algorithm reached to 73.5% after this experiment. To show the effect of negation on sentiment scores, we are multiplying the sentiment scores of words appended with NOT by -1.

TABLE IV. Results After Handling Negation

Product Name	Accuracy
Norton	73%
Canon G3	85%
Cannon S100	76.25%
Nokia 6600	75.70%
Hitachi Router	74.70%
IPod	76.56%
Average	76.86%

B. Valence Shifter

It is a very challenging task to reflect and distribute the effect of some negative words to its surrounding opinion words. Most sentiment analysis system perform well on majority of text classification problem, but a particular linguistic feature i.e. Valence shifter always poses challenges and problems to these systems. The study showed that almost 15% sentences in reviews contain valence shifters and handling them correctly significantly increases the classification accuracy [28]. Simple example of a review containing valence shifter is "This is not a good book," but not to our surprise, not many reviews are as straightforward as shown. Especially when consumer put a bad review about any product, they do not express their opinion very directly. For example: "The overly detailed approach makes it a hard book to recommend enthusiastically", although the reviewer is discouraging the readers to read the book but since the sentence contains words like enthusiastically and recommend, the overall sentiment score of the sentence might get positive score. In our proposed approach, after achieving 77% accuracy, we further enhance our analysis with Negation and run different experiments by changing the multiplication factor.

Previously, some model verbs if found in the opinion array receive their score from either Senti140 or CSUIC lexicon, and some of them were part of the Stop word list. We interchange their score, i.e., if previously some model verbs were getting their score from lexicon; we include those verbs in negative words list so that they could be treated as a negative word; hence, handled by the negation handler routine. If any model verb was part of the stop word list, we remove it and get its score from the lexicon. No model verb shows significant improvement over accuracy when handled individually by adding to the Negative word list or by removing from the Stop Word list. The reason for former is the number of words found in Senti140 lexicon. The dictionary is very large and almost contains the score of every word. Therefore, when we add these model verbs in negative word list, the accuracy does not increase. In contrast, some model verbs like should contribute best if ignore and included in Stop Word List.

C. Diminisher

Diminisher like valence shifter affects the score of preceding opinion words by some factor. Like negative words, diminishers also have considerable effect on sentiment analysis. In our experiment, we have tested some diminisher words and instead of using a multiplying factor of -1 like we have used for negative word list, we use Senti140 score to analyze the results. Following are the results of Diminisher words. The diminishers we used to analyze the changes were *below, few, over, small, down* etc.

TABLE V. RESULTS WHEN SCORE OF SENTI140 WAS USED FOR DIMINISHER WORDS

Product Name	Stop Word List	Using Senti140 dictionary
Norton	72%	73%
Canon G3	84%	83%
Cannon S100	76%	74%
Nokia 6600	77%	76%
Hitachi Router	73%	74%
IPod	77%	76%

V. ASPECT RANKING

After receiving the final sentiment score of top 20 aspects, we rank these aspects' sentiment score using bar graph. The X-axis defined the features or aspect, whereas the Y-axis defined the score. Each product has two separate bar graphs, one for positive, and one negative score aspects.

Figure 2and 3 shows the experimental resultsof **Cheetos.** In figure 2, our system has identified positive aspects of *value*, *chips*, and *package* and give score of 12, 0.9, and 0.3. Similarly, in figure 3 the negative aspect list contains an important aspect like *price* with score of -0.3. Since the product does not contain much objective reviews on Amazon.com, the systemfail to identify some good aspects. Moving further, we run our analysis on two specific smart phonei.e. **Samsung Galaxy S5 and IPhone 6.** The system identified that *fingerprint scanner*, *picture quality*, *battery life*,and *front camera* are the good aspects of S5 (shown in figure 4),whereas*apps*and*pixel density in pictures* are some bad aspects of it. In contrast, the IPhone 6 has *aluminum body*, *design*, and *elegant case* as positive aspect (shown in figure 6) with *button stabilizer* and *mute switch* as negative aspect (not shown in figure). Interpreting the results, customers who prefer better design and body of smart phone to its camera and battery can easily go for IPhone 6, whereas those who prefer good camera with longer battery life can buy Galaxy S5.

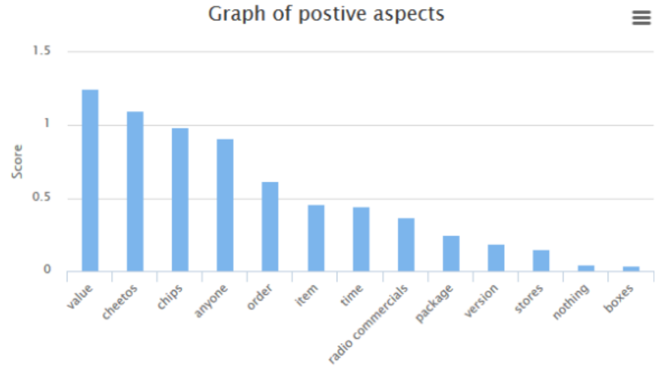


Fig. 2. Positive aspects of Cheetos shown in the form of bar graph

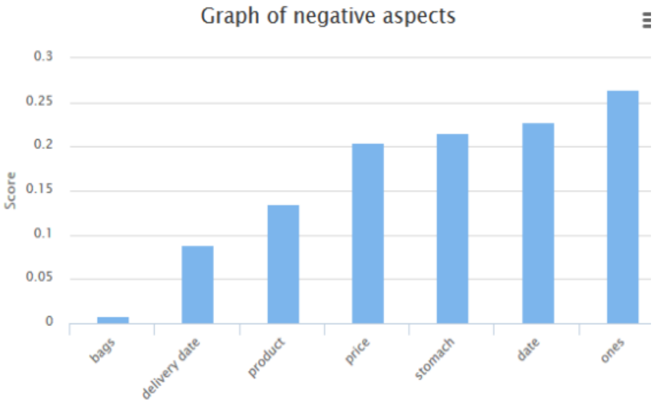


Fig. 3. Negative aspects of Cheetos shown in the form of bar graph

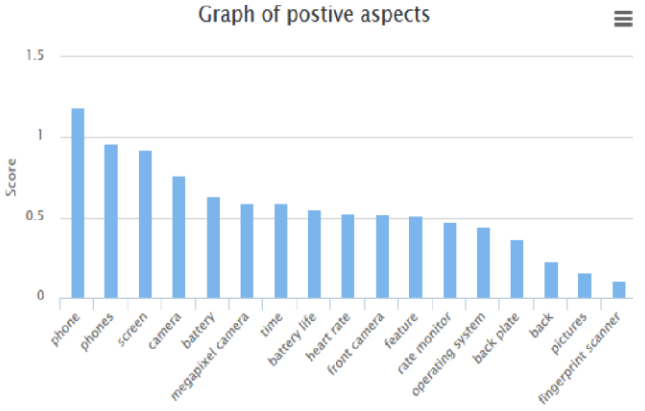


Fig. 4. Positive aspects of Samsung Galaxy S5 shown in the form of bar graph

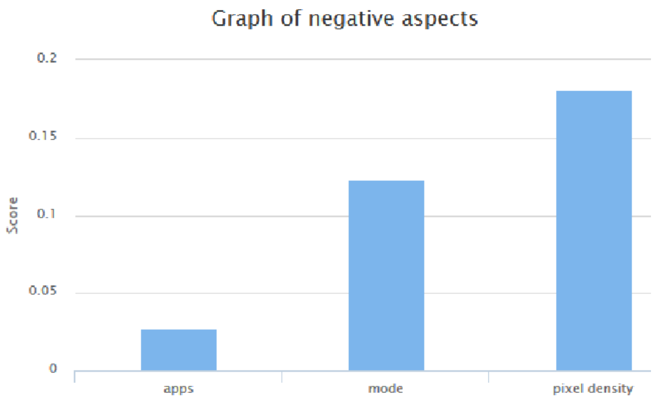


Fig. 5. Negative aspects of Samsung Galaxy S5 shown in the form of bar graph

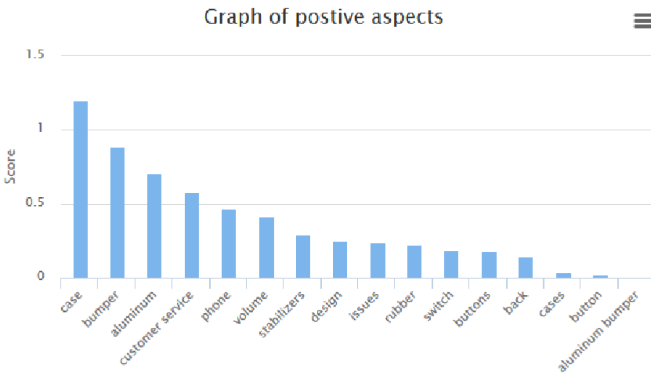


Fig. 6. Positive aspects of IPhone 6 shown in the form of bar graph

The results are dependent on the availability of not only reviews, but also some good reviews. These results are extracted from the reviews of Amazon only, and we are confident that if we increase the dimension of our search, the results will improve. In addition, the results of all the products either technical or non-technical category are sensitive to what reviews people are giving on Amazon. It is possible that many products do not have good or useful reviews on Amazon, and as a result, the result set may show meaningless aspects.

VI. Conclusion

In this research, we have studied and analyzed different factors that could affect the sentiment scores of opinion word. We have studied the importance of having relevant data for the reliability of any experiment and discussed the challenges of extracting target text from un-organized and un-structured spider of web pages. Further, we have discussed the importance of identifying aspects in the problem of product aspect ranking through sentiment analysis. We have proposed a lexicon based approach for product aspect ranking through sentiment analysis. We have exploited the impact of good lexicon on classification accuracy by experimenting with different short and rich lexicons. We also study the importance of handling negation in sentiment analysis task. For handling negation, we proposed a simple and effective approach of making a customized list of negative words and showed the increase in accuracy. Experimental results also proved that if a proper rich lexicon like **CSUIC** is used for product aspect ranking, then we do not need to handle the valence shifters (VS) and diminishers explicitly. The experimental corpus of review contained 5000 reviews of five different products. We have achieved 78% accuracy in finding the sentiment score of product aspects when used lexicons of CSUIC and Sentiment140 with windows size of 4.We also achieved 77% accuracy on restaurant training dataset of SemEval 2014 with 75% accuracy on Laptop Training dataset. Likewise, our system also achieved 75%accuracy on both trailing dataset of laptop and restaurant. We maintained a customized list of stopping words and negative words, which helped us to deal with negative words more accurately and without isolation.

In future, we will expand our domain from reviews and run our proposed approach in Social media text like Tweets and Facebook status. In addition, we will also experiment with other crude heuristics for identifying aspects from the text instead of targeting nouns only. Likewise, in sentiment analysis, we will expand our opinion dictionary with different other lexicons and analyze its final effect on our current approach. We will also work on neutral sentiment score of aspect and accumulate how to represent an aspect if it has a sentiment score of zero.

References

[1] Anonymous, “Larche Digital Media,” Larche, 2014. [Online]. Available: http://larchedigitalmedia.com/45-awesome-review-digital-marketing-stats-you-should-know/. [Accessed 12 May 2015].

[2] M. H. Bing Liu, “Mining and summarizing customer reviews,” in Knowledge discovery and data mining, ACM, Newyork, 2004.

[3] J. Y. M. W. T.-S. C. Zheng-Jun Zha, “Product Aspect Ranking and Its Application,” IEEE, vol. 26, no. 5, pp. 1211-1224, 2014.

[4] B. P. a. L. L. a. S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Empirical methods in natural language processing, Stroudsburg, 2002.

[5] S. J. Y. D. H. Hanhoon Kang, “Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis,” Expert System with Applications, vol. 39, no. 5, pp. 6000-6010, 2012.

[6] S. L. D. M. P. Kushal Dave, “Mining the peanut gallery: opinion extraction and semantic classification of product reviews,” in 12th international conference on World Wide Web , Pasadena, 2003.

[7] A. B. B. R. D. K. A. M. Michael Wiegand, “A survey on the role of negation in sentiment analysis,” in Negation and Speculation in Natural Language Processing , Uppsala, 2010.

[8] M. A. G. Diana Maynard, “Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis,” in Proceedings of LREC, 2014.

[9] E. Filatova, “Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing,” in European Language Resources Association (ELRA), Istanbul, 2012.

[10] A. Z. Livia Polanyi, “Contextual Valence Shifters,” in Computing Attitude and Affect in Text: Theory and Applications, vol. 20, California, Springer Netherlands, 2006, pp. 1-10.

[11] D. I. Alistair Kennedy, “Sentiment Classification of Movie Reviews Using Contextual Valence Shifters,” Computational Intelligence, vol. 22, no. 2, pp. 110-125, 2006.

[12] Z.-J. Zha, “Product Aspect Ranking and its Application,” IEEE "KNOWLEDGE AND DATA ENGINEERING", vol. 26, no. 5, pp. 1211-1224, 2014.

[13] H. a. Liu, “Opinion Mining, Sentiment Analysis, Opinion Extraction,” 15 May 2044. [Online]. Available: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. [Accessed 9 May 2015].

[14] D. Ashour, The Impact of Text Processing and Term Weighting on Arabic Text Classification, Gaza: The Islamic University, 2010.

[15] D. Torunoglu, Analysis of Preprocessing Methods on Classification of Turkish Texts, Turkey: Dogus University.

[16] P. Q. Teresa Gonc ̧alves, “Evaluating preprocessing techniques in a Text Classification PRoblem,” in International Conference, Portugal , 2005.

[17] N. Z. a. J. Liu, “The Mining Mart Approach to KNowledge Discovery in Database,” Intelligent Technologies for Information Analysis, pp. 47-65, 2004.

[18] R. A. V. Srividhya, “Evaluating Preprocessing Techniques in Text Categorization,” International Journal of Computer Science and Application, pp. 47-11, 2010.

[19] S. G. Alper Kursat Uysal, “The impact of preprocessing on text classification,” Information Processig and Management, vol. 50, no. 1, pp. 104-112, 2014.

[20] L.-F. C. H.-J. L. Wen-Hsiang Lu, “Anchor Text Mining for Translation of Web Queries: A transitive translation approach,” ACM Transactions on Information Systems , vol. 22, no. 2, pp. 242-269, 2004.

[21] Y. C. C. C. J. T. H. Myle Ott, “Finding deceptive opinion spam by any stretch of imagination,” in Human Language Technologies, Stroudsburg, 2011.

[22] B. Liu, “Sentiment Analysis Layers,” in Sentiment Analysis and Opinion Mining, Claypoool Publisher, 2012, pp. 10-12.

[23] S. Muhammad, "Saif Muhammad Homepage," 07 July 2015. [Online]. Available: http://www.saifmohammad.com/WebPages/ResearchInterests.html. [Accessed 9 August 2015].

[24] B. L. Nitin Jindal, “Opinion Spam Analysis,” Chicago, Illinoise, 2013.

[25] S. Evaluation, "Metashare," 9 December 2014. [Online]. Available: http://metashare.ilsp.gr:8080/repository/search/?q=semeval. [Accessed 9 August 2015].

[26] E. Lapponi, “Representing and Resolving Negation for Sentiment Analysis,” in Data Mining Workshops (ICDMW), Brussels , 2012.

[27] C. Y. W. M. Lifeng Jia, “The effect of negation on sentiment analysis and retrieval effectiveness,” in 18th ACM conference on Information and knowledge management, New York, 2009.

[28] L. Dillard, “Sentiment Classification,” University of Washington, Washington, 2007.

Adoption of Biometric Fingerprint Identification as an Accessible, Secured form of ATM Transaction Authentication

Michael Mireku Kwakye
Faculty of Informatics
Ghana Technology University College
Accra, Ghana

Hanan Yaro Boforo
Genkey Solutions Africa
Accra, Ghana

Eugene Louis Badzongoly
Faculty of Informatics
Ghana Technology, University College
Accra, Ghana

***Abstract*—Security is continuously an important concern for most Information Technology-related industries, especially the banking industry. The banking industry is concerned with protecting and securing the privacy and data of their customers, as well as their transactions. The adoption of biometric technology as a means of identifying and authenticating individuals has been proposed as one of the varied solutions to many of the security challenges faced by the banking industry. In this paper, the authors address the ATM transaction authentication problem of banking transactions using fingerprint identification as one form of biometric authentication. The novel methodology adopted proposes the use of an online off-card fingerprint verification, which involves the matching of live fingerprint (templates) with pre-stored templates read from the ATM smart card. The experimental evaluation of the proposed methodology presents a system that offers a faster and relatively better security of authentication, as compared to previous and existing methodologies. Moreover, the use of BioHASH templates ensures an irreversible cryptographic hash function, facilitates a faster authentication, and enables an efficient framework of detecting potential duplicates of banking account holders.**

Keywords—Information Technology; Automatic Teller Machine; Biometrics; Fingerprint; BioHASH; Token

I. Introduction

The revolution of Information Technology (IT) has generated a lot of development and innovation in the areas of business, academic and industrial research, and healthcare, amongst others. Technology has become the backbone of every organization and an appreciable volume of system, human, and financial resources are utilized in the adoption, development, and incorporation of these technologies into the day-to-day activities of an organization.

The emergence of the internet has paralleled the IT revolution and facilitated IT development into various innovations [1]. The internet has changed the manner in which individuals and organizations interact and communicate with each other [2]. The internet has also changed the way businesses operate, and as a result the introduction of electronic commerce has enabled easy, accessible, and efficient medium for businesses to effectively interact with their customers and partners all over the world [2].

The banking industry has become one strategic industry that utilizes Electronic Commerce (E-Commerce) [3]. The past decade has seen an increase in the adoption of technological innovation in the banking sector. The increase is mainly being driven by the desire of the banks to remain profitable and competitive [4]. Automatic Teller Machines (ATMs), telephone banking, and online banking make up more than 50% of the banking transactions in some developed economies, like the United States of America (USA); and this is growing at a rate of 15% annually [2]. Electronic banking presents the banking industry with an electronic and remote distribution channel, which serves as an electronic market place where consumer with individuals and business can conduct their financial transactions virtually [4].

With the increase in electronic banking, one major concern for this medium of transaction processing is security and privacy [2]. Most internet users get worried about privacy issues, including transparency in the collection, use, and disclosure of their personal information. A relative number of users are also worried about the security of their bank accounts and transaction details [2]. Electronic banking comes with a high level of exposure to common cyber-related risks. Varied risks, such as, information hacking, cyber-sabotage, and cyber-terrorism, amongst others, all together adopt unique ways of attacking a system [5]. Electronic banking requires the implementation of high-quality security features and procedures [2]. One of such security features is the use of biometrics in identifying and authenticating an individual user to a system. Biometric technologies enable the identification and authentication of an individual user based on the physiological and/or behavioural characteristics [5]. Though biometric systems have been successfully adopted and deployed in areas, such as, criminology, health, electioneering procedures, and immigration control, there is little research and implementation pertaining to the banking industry [5].

In this paper, the authors introduce a framework that offers a viable biometric technology implementation in the banking industry. The primary focus is the adoption of fingerprint identification as a biometric measure for an accessible and secured form of ATM and card technologies security in the area of electronic banking.

The motivation of the authors is to employ the concept of BioHASH templates, which ensures an irreversible cryptographic hash function, facilitates a faster authentication, and enables an efficient framework of detecting potential duplicates of banking account holders. The authors' key contribution in this paper is the adoption of an online off-card fingerprint verification, which involves the matching of live fingerprint (templates) with pre-stored templates read from the ATM smart card.

The technical contributions are summarized, as follows;

- The authors design a biometric enrolment system that requires new customers of a financial institution to register their biometric information together with their biographic information during account opening;
- The authors propose the design and implementation of an online off-card verification and biometric authentication system on ATMs that works without a remote connection to an application server for verification and authentication on the ATM system;
- The authors employ the BioHASH template technology for a cryptographic hash function in the identification, verification, and faster matching of biographic data.

The rest of the paper is organized as follows. In Section II, the authors review the fundamental background studies on ATM banking and biometric authentication. In Section III, the authors discuss the proposed biometric (fingerprint) methodology framework. Here, the authors address the overview of the proposed system for the adopted methodology approach. In Section IV, the authors address the proposed system architecture, discuss the modules encompassing the proposed architecture, and outline the overall system operation and flowchart. In Section V, the authors address the propositions of the fingerprint methodology; where the authors outline the merits for the accountholder and transaction processing and authentication procedures. In Section VI, the authors address the implementation, testing, and evaluation of the methodology framework; as an effective approach in providing security in ATM systems. The authors discuss the related work and comparison of other approaches in Section VII. Finally, in Section VIII the authors conclude, discuss open issues and the areas of future work.

II. Background

Information Technology (IT) has brought about improved efficiency and effectiveness in the operations of most organizations. This trend has posited an assertion that currently IT is the backbone of every organization. Electronic Banking (E-Banking) is the provision of banking products and services through electronic delivery channels. Services offered via electronic banking channels include, Automated Teller Machines (ATM), Internet banking, and Mobile banking, amongst others. ATMs have existed in recent past and are found in most parts of the world for different forms of electronic transactions and processing. ATMs have become the most visible pieces of electronic hardware in the banking sector, and they are also the fastest growing element in banking. ATMs became popular more than 20 years and as a result banks and their respective client users have since gained a lot of advantages from the use of ATMs [6].

Biometric authentication using fingerprint identification is seen by many as the solution to most of the theft and fraud cases being reported in the use of ATM systems and ATM cards. Biometrics-based authentication offers several advantages over other authentication methods, as there has been a significant surge in the use of biometrics for user authentication in recent years [7]. Onyesolu and Ezeani (2012) [8] in their study found that, majority of their respondents chose fingerprint identification as the preferred biometric identification solution to ATM card theft and fraud. In the proposed biometric-based ATM authentication system designed and developed by the authors in Oko and Oruh (2012) [7], the result of their methodology and testing evaluated that biometric authentication on ATM systems was practicable and could be implemented in production environments.

Daula and Murthy (2012) [9] developed an embedded fingerprint identification system which is used for ATM authentication. Their system makes use of GSM modem for authenticating users. The system required banking institutions to capture the biometric fingerprints and cellular (mobile) number of customers during account opening. At the ATM system console, the customer of the bank places his/her fingers on the fingerprint scanner attached to the ATM machine. The system on the ATM then compares the fingerprints to the previously captured fingerprints. If the fingerprints are found to be a match, a 4-digit code is generated and sent to the customer mobile phone. These 4 digits are then entered on the ATM. This system does not require the use of an ATM card. The system is secured because it securely verifies and authenticates the identity of a cardholder who tries to do transaction through the ATM.

Biswa et al. (2012) [10] also conducted a research which was aimed at developing a crypto-bio authentication system in ATM banking systems. Their system relied solely on the usage of retinal image. Hossian et al. (2013) [11] proposed a biometric authentication scheme for ATMs, their system made use of an Advance Encryption Standard (AES) processor instead of the Triple Data Encryption Standard (3DES). The study concluded that, the usage of an AES processor and fingerprint biometric identification made the ATM transaction more secured.

Previous works in the area of biometric fingerprint authentication of ATM cardholder followed a client-server paradigm. The ATM system captures the scanned fingerprints of a cardholder who wants to perform a transaction and transfers it to a biometric verification application on a remote server. The application connects to the biometric repository (biometric database) of the financial institution that owns the ATM system, and verifies the submitted fingerprint templates against the pre-existing templates of the cardholder in the biometric repository. The system developed by Daula and Murthy (2012) [9] follows the same paradigm and improves the procedure in generating a 4-digit Personal Identification Number (PIN) that is sent to the cardholder via a Short Messaging Service (SMS). The 4-digit code is then entered and verified on the ATM.

Venkatraman and Delpachitra (2008) [5] in their study concluded that, though biometric has been successfully deployed in areas, such as, immigration control and criminology, there is little literature on their implementation in the banking sector. Their study identified 4 main categories of issues that are critical to the viable adoption of biometric-based authentication in New Zealand. These factors are listed as; Technological, Management, Legal and ethical, and Monetary.

The biometric identification systems described above have some flaws that can impact on their performance. The solution proposed by Daula and Murthy (2012) [9] relies on the use of a 4-digit PIN that is sent to the account owner via a SMS. This PIN is entered at the ATM terminal in order to complete the authentication process. The major flaw with this system is the delivery of an SMS is not 100% reliable because of the unreliable channels in telecommunication networks. There are instances where the SMS will fail to reach its destination or it could take quite a while in reaching its destination. Additionally, the system only limits the use of the ATM in connection to the ownership of a mobile phone.

Other systems as one described by the authors in Gelb and Decker (2011) [12] requires access to a central biometric database in performing identification or verification. However, this method can result in higher error rates depending on the number of templates being accessed. The system also poses privacy concerns because the biometric of an individual are stored in a central database [13]. Firstly, access to a central database during an authentication process can be quite slow depending of the number of templates stored on the database. Secondly, the use of a central database means that, authentication at the ATMs can only be done for people who have their biometric information stored on the central database that is being access by the ATM. This makes it impossible for a cardholder to perform transactions on ATMs owned by different financial institution even if they are using the same biometric vendors as the parent institution of the cardholder. This is mainly because their ATMs will most likely connect to a different biometric repository.

The proposed solution in this paper described in Sections III, IV, and V addresses the major flaws identified in the previous systems above, and will employ the use of smart card (debit cards) with the biometric information of the cardholder encrypted on the card.

III. Fingerprint Methodology Framework

In this Section, the authors address the problem statement leading to this research and propose a methodology solution that is efficient and secured enough in comparison to earlier and existing approaches.

A. Problem Definition

The continuous growth in the various paradigms of financial services, such as, Electronic Commerce, Internet Banking, and ATM banking requires the development and implementation of sound security systems and procedures [2]. ATM transactions require the designing and implementation of authentication mechanisms in a remote environment. The current systems of authenticating ATM transactions involve the use of an ATM card and a Personal Identification Number (PIN).

The major concern with this type of authentication is that, ATM cards can be cloned on one hand, and on the other hand, PINs are often shared with family relatives or close associates. Sharing of PINs happens when a cardholder (account owner) decides to allow a friend, an associate, or a family relative undertake some ATM transaction on his or her behalf.

ATM fraud is a major issue being faced by most financial institutions, research has shown that, there is a continuous rise on the number of ATM fraud being reported yearly [8].

B. Overview of Proposed Biometric ATM Methodology

The authors approach for designing an efficient biometric ATM solution for banking transaction is such that, an individual’s biometric details will be captured when opening an account. This biometric information will be sent to a third party biometric vendor (say, Genkey Solutions), and the vendor will process the biometric fingerprint information and generate BioHASH tokens from them. The BioHASH token is written onto the microchip of a smart card or a debit card. The captured biometrics are discarded after BioHASH tokens have been generated from them. At the point of authentication or performing a transaction at the ATM, the pre-stored biometric details on the smartcard and the live fingerprints captured at the ATM are sent to an authentication server in a secured manner. This approach ensures that verification is not done against a set of templates stored in a database. This decreases the possibility of a false reject identification and also gives the user more privacy; having his or her biometric information in his or her own custody [13].

The authors address the methodology using the diagram in *Figure 1* below. This system architecture outlines various components; such as, Registration Client, Biometric Database to store biometric templates, Biographic Database to store the biographic details of the account holder or customer, and the Biometric Vendors De-duplication Server. Communications between these various components is done over a secured network. Each of these components is described in Section IV (Proposed System Architecture).

IV. Proposed System Architecture

The proposed system architecture is modelled into 2 main modules; namely, the Registration Client module and the Biometric Authentication module. The authors discuss each of these modules, as follows:

A. Registration Client Module

The Registration System Module is split into 2 subsystems; namely, the Registration Client, and the Biometric Vendor’s de-duplication server. The Registration Client is made up of the following components: the Application Server, the Work Station (for capturing biographic, biometric details of applicants and also for generation and producing ATM cards), and the Storage (for customer biographic information and biometric information). The Biometric Vendor’s de-duplication server is made up of the following components; the REST Application Server, and the Storage (for biometric information).

The authors illustrate the diagrammatic description of this module in *Figure 2*. The diagram depicts the processing that takes place on each of the various components that make up the registration module. The Enrolment Workstation illustrated in the diagram is used for capturing both Biographic and Biometric data from a Customer. The biographic details are validated and stored in the Biographic Database, the Biometric data is sent to the de-duplication server (Biometric Web Server) for de-duplication to take place. The de-duplication server processes the fingerprints and generates BioHASH templates and matches those templates against templates previously stored in the Biometric Database.

B. Biometric Authentication Module

This system is operational on the ATMs when users are about to perform transactions. *Figure 3* below illustrates the system architecture for the biometric authentication module. The biometric authentication module is made up of the following components; the ATM Console, the Smart Card, and the Fingerprint Scanner (attached to the ATM console).

The system architecture (*Figure 3*) depicts the general flow of an online verification process. Every activity related to the verification takes place in the biometric authentication server. The ATM performs the following functions outlined below, before sending biometric information to the biometric authentication server (verification server), for verification and/or authentication. These are;

- Validate ATM card;
- Read the ATM card to retrieve biographic and biometric information (BioHASH templates) of the cardholder;
- Invoke Fingerprint scanner to take fingerprints of cardholder;
- Encrypt scanned fingerprints and information read from ATM card;
- Transmit biometric information to biometric authentication server for verification (authentication);
- Grant or deny access depending on the success or failure result, respectively, from the verification (authentication) process.

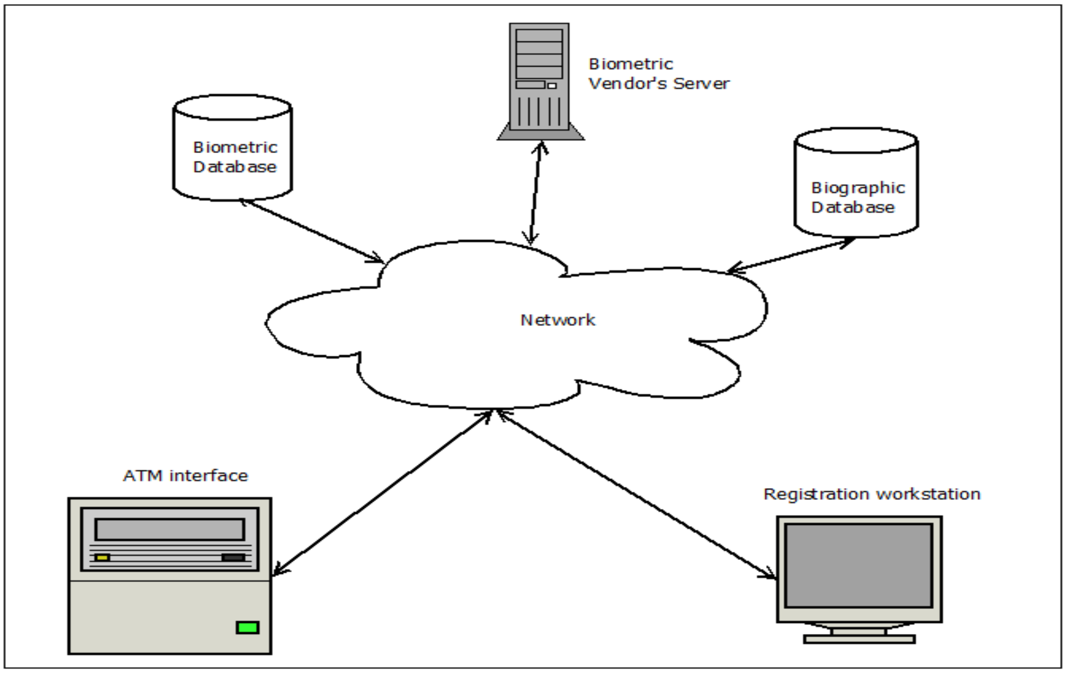


Fig. 1. Overview of Proposed System Design

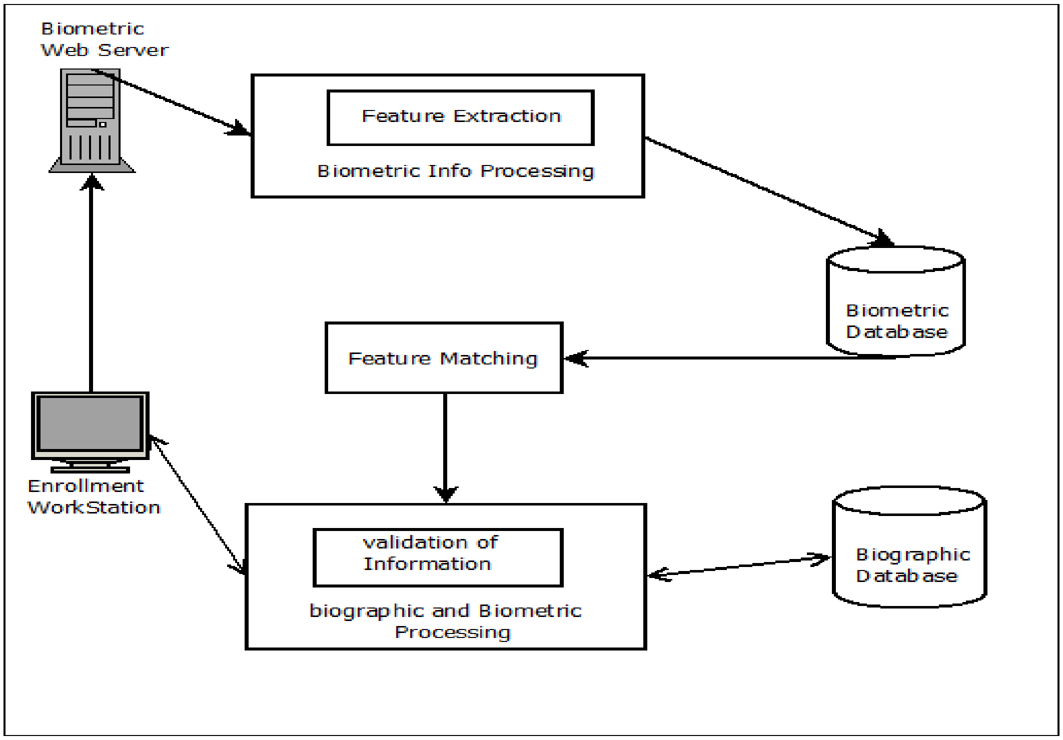


Fig. 2. Overview of the Registration Client Module

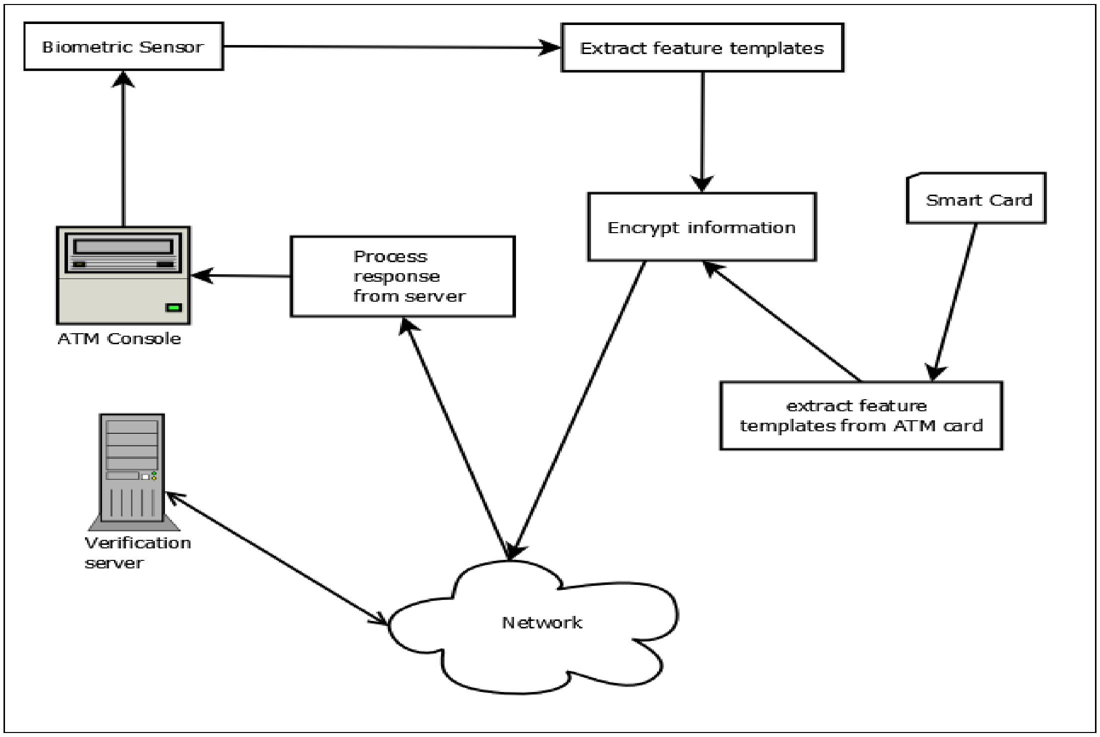

Biometric Sensor
Extract feature templates
Smart Card
Encrypt information
Process response from server
ATM Console
extract feature templates from ATM card
Verification server
Network

Fig. 3. Overview of an Online Biometric Authentication

C. Proposed System Operation

The proposed system will use smart cards with fingerprint validation instead of the usual PIN. The aim of this systemic approach is to address the defects that have been identified with current implementations of using PINs.

The system works as follows:

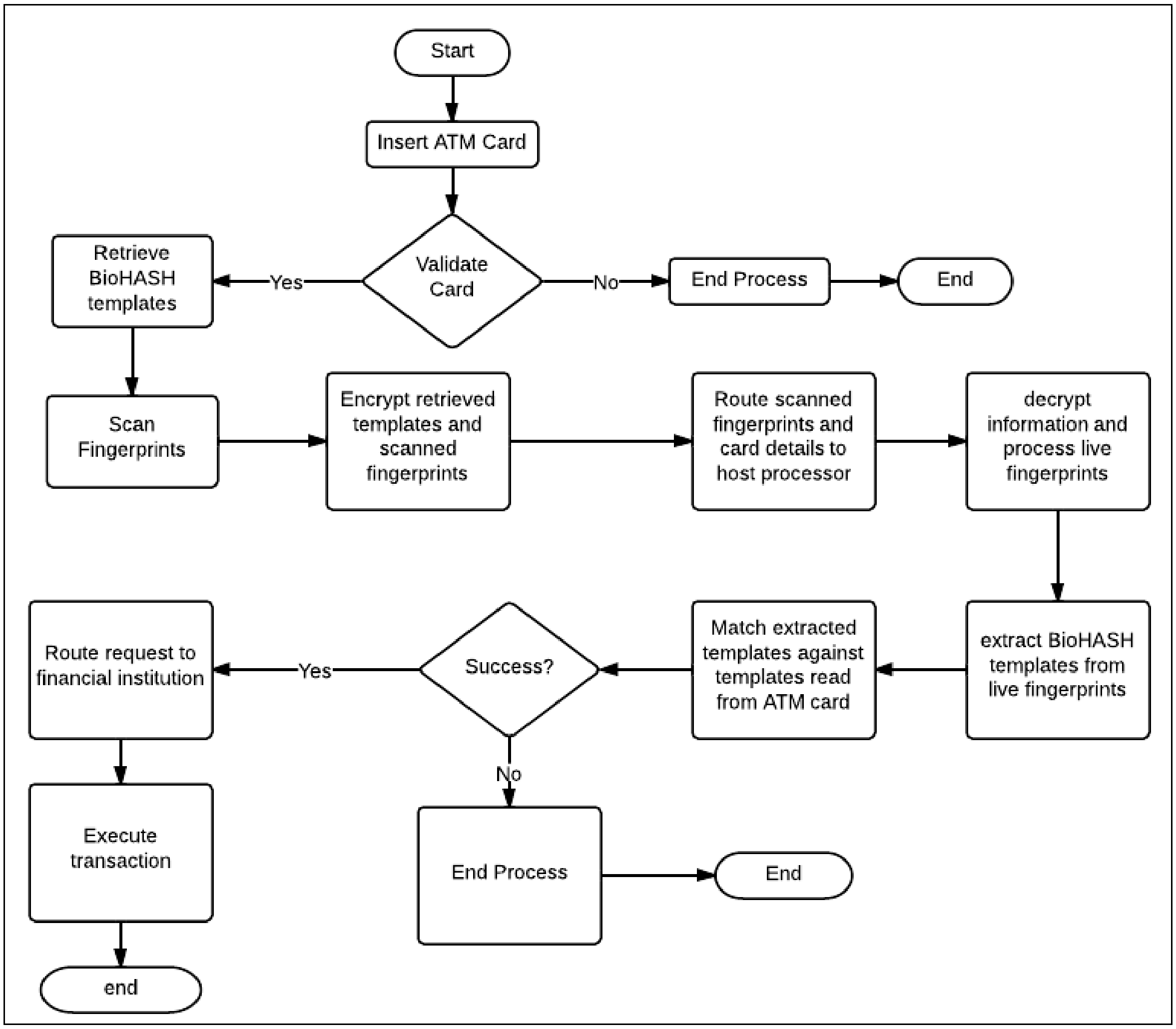

Start
Insert ATM Card
Validate Card
Yes
No
End Process
End
Retrieve BioHASH templates
Scan Fingerprints
Encrypt retrieved templates and scanned fingerprints
Route scanned fingerprints and card details to host processor
decrypt information and process live fingerprints
extract BioHASH templates from live fingerprints
Match extracted templates against templates read from ATM card
Success?
Yes
No
Route request to financial institution
Execute transaction
end
End Process
End

Fig. 4. System Flowchart of Online Off-card ATM Biometric Authentication

```
Algorithm 1:   System Flow of Online Off-card ATM Biometric Authentication
(1)  Insert card into ATM console.
(2)  ATM's processor validates card. Upon successful validation go to Step 4; failure validation go to Step 3.
(3)  ATM's processor ends process and ejects card.
(4)  ATM's processor retrieves information stored on the card.
(5)  Cardholder is prompted to scan his or her fingerprints.
(6)  Scanned fingerprints are encrypted.
(7)  Biometric information are retrieved from card and scanned fingerprints are transmitted to a host processor (verification server).
(8)  Verification server decrypts the information transmitted to it.
(9)  Verification server processes scanned "live" fingerprints and generate BioHASH templates from it.
(10) Server then matches the newly generated BioHASH templates against the BioHASH templates retrieved from the smart card (ATM card).
(11) Upon successful matching go to Step 13, otherwise a match failure means go to Step 12.
(12) Failure processing is reported back to ATM system and process is brought to an end.
(13) Verification server sends match success response to the concerned financial institution.
(14) The cardholder is then allowed to perform his or her financial transaction.
```

A cardholder inserts his or her card into an ATM console, the card is validated, and both biographic and biometric (BioHASH templates) details stored on that card are read. Once the information is read, the cardholder will be required to scan his or her fingerprints. After the scanning of the required fingerpints, the ATM's processor will encrypt the biometric details read from the card and the live fingerprints. This information is then transmitted to host processor (authentication server) via a secured network. At this point, the encrypted information is decrypted. The Verification server then processes the live fingerprints and generates BioHASH templates from them. The extracted templates are then matched against the BioHASH templates retrieved from the smart card.

Once matching is successful, the host processor (verification server) routes the client's request to the concerned financial institution or bank. The financial institution then validates that the supposed cardholder is allowed to successfully perform financial transactions.

Figure 4 illustrates the flow of activities of the proposed authentication system. The system seeks to make two changes to the current system of authentication on ATMs. The first change it seeks to make is to replace the use of a PIN with the use of a fingerprint scan. The second change is to drift away from authentication using a database of biometric information to authenticating users using biometric details that have been captured and written to their ATM cards. The functionality of the proposed system is explained by the steps outlined below (*Algorithm 1*):

V. Propositions of Biometric Fingerprint Methodology Framework

The authors' adoption of this proposed solution offers a number of advantages over previous and existing approaches, as well as offering an efficient and security-aware solution for current ATM transaction processing. The authors explain below the propositions of merits for the novel methodology of biometric fingerprint authentication.

1) Flexibility: The proposed system offered a more flexible design where the user does not need to apply a PIN code alongside using the biometric fingerprint. Moreover, all the biometric and biographic information are stored on the smart card. This enables easy transaction on ATM consoles and the cardholder does not need to remember PINs and passwords. This functionality is a major advantage over previous approaches where a card holder combines a PIN code alongside his biometric fingerprint for transaction authentication.

2) Scalability: Scalability is a major concern when developing web applications. In this regard, the proposed system offered architectural design and interfaces where multiple authentications and transactions are performed concurrently with less data flow traffic. Here, the user access at any point in time could range even to a 1000 persons. Additionally, the platform supports the design of related scalable applications. In comparison to other approaches, these systems rather use desktop applications which limit concurrency usage of the application and/or builds up communication traffic during transaction authentication and processing.

3) Fast User Authentication: The proposed system design enables an authentication process that is more efficient and fast enough, in comparison to existing system approaches. This functionality is achieved because of the transfer of smaller file sizes of BioHASH templates over communication networks. Additionally, the matching procedures of the "live" and pre-stored BioHASH templates are easily adjudicated because biometric database needs to validate if tokens from both templates are the same. There is no need to do a match among a lot of pre-stored biometric data, with a resultant effect of high false reject errors. Comparing the use of BioHASH templates to other approaches, this functionality is a major merit. Firstly, the procedures do not require the transfer of the entire biometric data, but rather the template tokens; which are much smaller in file sizes. Secondly, there is no need to match pre-stored biometric templates.

4) Privacy Preservation: The privacy of account holders and their transactions are ensured in this proposed system. This is achieved because of the feature of both the encrypted biometric and biographic information of account holders are stored on the smart card; leaving only encrypted token BioHASH templates on the vendor's biometric database. This feature is quite beneficial because no biometric and biographic data is stored anywhere, whether on the vendor's database.

5) Efficient Security: The system that was designed offered a better and a more efficient system of authentication than the existing systems. The use of BioHASH templates ensures an irreversible cryptographic hash function, and also ensures that the original biometric are discarded as soon as the BioHASH value is derived. This means that the scanned biometrics are never stored or used in the matching process. Since matching is not done against templates stored in a database but rather against the BioHASH templates read from the ATM card, the rate of occurrence for both false acceptance and false rejections are drastically reduced. This makes the system more efficient and more secured. Furthermore, this system prevents the cloning of ATM cards which are prevalent with other approaches.

VI. Implementation and Evaluation

In this section, the authors discuss the implementation, testing, and evaluation work based on the proposed system methodology. The authors present the implementation framework and the procedures, and they discuss and analyze the evaluation results.

A. Implementation

The authors describe the implementation framework of various techniques and processes needed in delivering a secured system of ATM transaction processing. This sub-section focuses on the experimental setup and database design, the development environment deployed, the implementation testing applied, as well as the varied evaluations assessments to ascertain the efficiency of the proposed ATM transaction authentication methodology addressed in Sections III.B, IV.A, IV.B, and IV.C.

1) Experimental Setup and Database Design: The authors implemented the design using various sub-modules as part of system development; namely, Registration Client, De-duplication Server, and Set of Databases (Biographic and Biometric). The Registration Client is made of the following components;

- User Interface;
- Various Entities (Classes);
- Adjudication Client;

- Card Generation Client.

The Registration Client is supported by the Biographic database. There is the biometric de-duplication server and the biometric database used in storing generated BioHASH templates. The database was implemented using 2 different databases; as follows; Biographic, Biometric. The Biographic database is used solely by the registration module to store the biographic and account details of a particular account cardholder. The Biometric database is used by the third party biometric provider to store the biometric data for the customers of the bank.

2) Programming and Code Generation: The Registration Client's interface was developed as User Interface Form. Entities, such as, Staff, Customers, Branches, Account, AccountTypes had individual user interface forms that were designed to either create or edit them. Moreover, classes were developed to handle the business logic of each of the entities listed above. The classes created were used for basic operations like retrieval of Customer, Branch and Account information. The classes developed also handled the insertion of new records into the respective database tables; as well as deleting and updating of information concerning the various entities. A data access class was also created to handle connections to the database. In summary, 9,978 lines of code were written for the entity classes in support of the business logic of the system. In coding the User Interfaces, 8,526 lines of code were written. The de-duplication request and response classes were made up of 10,132 lines of code.

3) Integrated Development Environment (IDE): Microsoft Visual Studio was used to develop stand-alone, web applications, web sites and web services. The programming languages used was C#, which is well-integrated with the .NET platform.

4) Futronic Fingerprint Scanner and SDK: Futronic's FS80 USB 2.0 fingerprint scanner was chosen because of the extensive support it has for a number of platforms, such as, Windows and Linux. The device was also chosen because it has Software Development Kit (SDK) support for both Java and .NET platforms. Futronic fingerprint scanners are very durable and they also use advanced CMOS sensor technology which helps in delivering very high quality fingerprint images. Additionally, they are very fast in capturing fingerprint images.

5) Card Printer: Zebra ZXP series 1 card printer was used as the ideal printer for the proposed system. The card printer provides high quality card printing.

6) Smart Card Reader/Writer: The proposed system is fitted with a smart card reader and writer, for the dual usage of the card envisioned (to be used by both ATMs and POS devices). A supposed reader/writer should read both contactless smart cards, and virtually any other type of smart card. For this purpose, the OMNIKEY 5321 Smart Card Reader/Writer was chosen. This reader/writer offers a dual interface that allows for the use of both contactless and contact smart cards.

B. System Testing

The authors performed a number of tests to ascertain the effectiveness and efficiency of the system that was developed. The authors explain below these set of tests for the modules adopted in the system methodology.

1) Unit and System Testing: This form of testing was done in 3 stages; the first stage of testing focused on each of the entities. In this regard, unit test were written to validate the data that were captured for entities, such as, Staff and Customers. The second stage focused on testing the Registration Client independently; the test involved the entire information flow of creating and editing the details of the various entities. The reason for this is to ensure that there is a seamless information flow from one point in the Registration Client to another. The last stage of testing focused on the interaction between the REST server and the registration module. This test also included the card generation and printing.

2) User Interface Testing: The testing procedure on the User Interface for the Registration Client was done in 2 ways. These tests involved the design of the User Interfaces. The first process involved the Cognitive Walkthrough approach. Users were given a series of tasks to perform on the Registration Client, and the feedbacks collated were used to further refine the design of the User Interface. The second process adopted identified inconsistencies in the design of the User Interface. This test focused on the appearance of the Interface and not its functionality. This test focused on the font sizes used, colour, terminologies and layout.

3) Registration Client Testing: The Registration Client was also tested using the Cognitive Walkthrough approach. These tests involved the enrolment procedures that have to be followed in registering a Customer. Users were given the task to enrol a Customer, the feedback received from them during each stage of the enrolment were later used to refine the design and functionality of the Registration Client.

C. Evaluation

The authors assessed the functionalities of the proposed system based on various metrics and discussed the merits over previous system approaches. Moreover, the authors quantitatively analyzed the set of procedures (and sub-procedures) involved in the overall system framework and methodology.

The authors present below in TABLE I. the average response time for the set of procedures in the overall system framework and methodology. Here, the authors analyze the procedures of acquisition of fingerprint using the scanner, completion of the enrolment process for registration, and the online off-card verification during a cardholder transaction authentication. The collation of response times were based on 10 successive attempts of careful system testing for each of the sub-procedures per each procedure.

VII. Comparison to Other Approaches

There have been a few literature and studies in the area of biometric fingerprint authentication on ATM systems. Though the studies explain varied methodologies and techniques and present significant contributions, some pertinent problems are not addressed in-depthly or still unresolved. In this section, the authors discuss these approaches and comparatively explain how the proposed methodology performs better.

A. *Biometric ATM Authentication against Central Database (Tokenless Authentication)*

This form of biometric authentication does not require the use of an ATM card (Token). This system is currently being implemented in rural areas in India [12]. The system requires access to a central biometric database in performing as part of identification or verification. Access to a central database during an authentication process can be quite slow depending of the number of templates stored on the database.

TABLE I. Quantitative Summary of Average Response Time for System Procedures

Procedure	**Average Response Time (s)**	**Sub-procedure**	**Average Response Time (s)**	**Comment**
Acquisition of Fingerprint using Scanner	32.40	Not Applicable	Not Applicable	The adopted fingerprint scanner (Futronic FS80 USB 2.0) can only scan a single finger at a point in time. An ideal scanner, a Slap Fingerprint Scanner, will scan 4 fingers at the same time and that will appreciably reduce the response time.
Completion of Enrolment Process	380.00	Capture Biographic Information	150.000	The Enrolment process involves the capturing of both biometric (Fingerprints and Photograph) and biographic data. The time taken to complete this process is influence by the typing speed, the ease of taking the applicant’s photograph, the fingerprint acquisition process, the network latency, and database performance.
		Capture Biometric Information (Photograph)	78.000	
		Capture Biometric Information (Fingerprints)	120.000	
		Biometric De-duplication Process	132.000	
Online Off-card Verification	5.55	Capture “Live” Fingerprint	3.200	This procedure generally affected by network latency from the ATM console to the vendor’s (biometric) database. Furthermore, the matching process is impacted upon by the efficiency and speed in information processing on the vendor’s database.
		Read Biometric Information from Smart card	0.800	
		Extract Template from “Live” Fingerprints	0.500	
		Matching of Templates	0.045	
		Recording of Response and Request	1.000	

A database read could become the bottle neck in the system and this might lead to a large number of people abandoning the use of biometric ATMs.

Moreover, the use of a central database means that authentication at the ATMs can only be done for people who have their biometric information stored on the central database that is being access by the ATM. This makes it impracticable for a cardholder to perform transactions on ATMs owned by different financial institution even if they are using the same biometric vendors as the parent institution of the cardholder. This is mainly because their ATMs will most likely connect to a different biometric repository. The benefit of this system is that it reduces the cost of having to issue ATM cards and other related costs. *Figure 5* illustrates a general architecture of the system. The major drawback of this system of authentication is that, matching the biometric details an account holder against a large database of biometric details can lead to high false reject errors [13].

B. *Biometric ATM Authentication Using GSM Modem*

Daula and Murthy (2012) [9] in their work identified some flaws with the traditional ATM systems authentication. They introduce a solution that was aimed at addressing the flaws in the traditional system. *Figure 6* illustrates an overview of the system flow of their solution. The system flowchart only depicts how a User interacts with their solution and ignores the interactions of a system administrator. The system requires the cellular (mobile) number and fingerprint of an account holder. The account holder scans his or her fingerprint at the ATM. These fingerprints are authenticated and an Short Messaging Service (SMS) with a password is sent to the mobile phone of the account holder. The account holder is then required to enter the password sent, via SMS, to him or her to complete the authentication process.

The system explained above has major technical flaws, as discussed below:

- The system relies on the safe and timely delivery of an SMS that contains a password to be used for the final step of validation at the ATM. The major issue with this is that, SMS deliveries are not always reliable. It also means accountholders who do not have their mobile phone with them cannot perform financial transaction via the ATM.

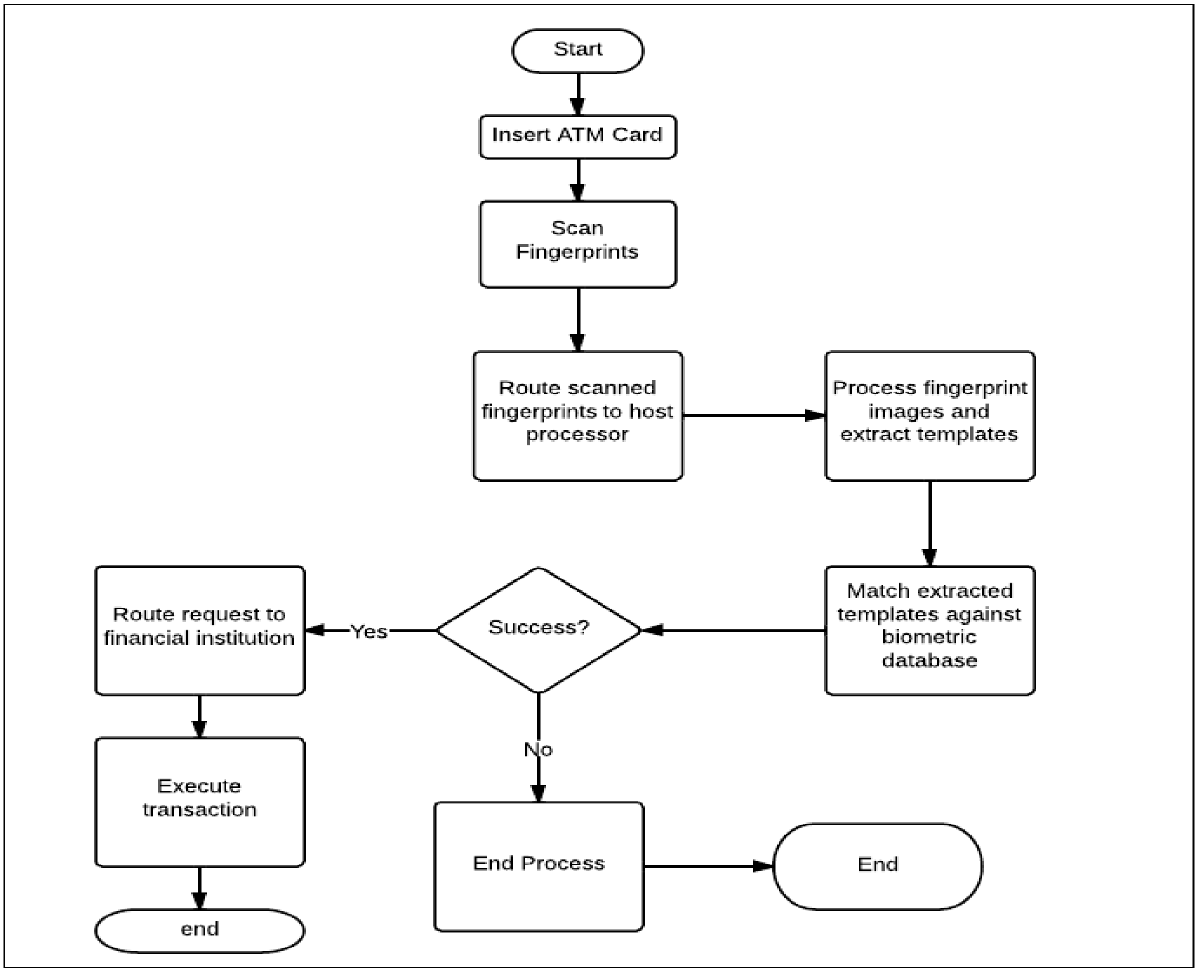

Start
Insert ATM Card
Scan Fingerprints
Route scanned fingerprints to host processor
Process fingerprint images and extract templates
Match extracted templates against biometric database
Success?
Yes
Route request to financial institution
Execute transaction
end
No
End Process
End

Fig. 5. Flowchart of Proposed System of Tokenless Biometric Authentication, Gelb and Decker (2011)

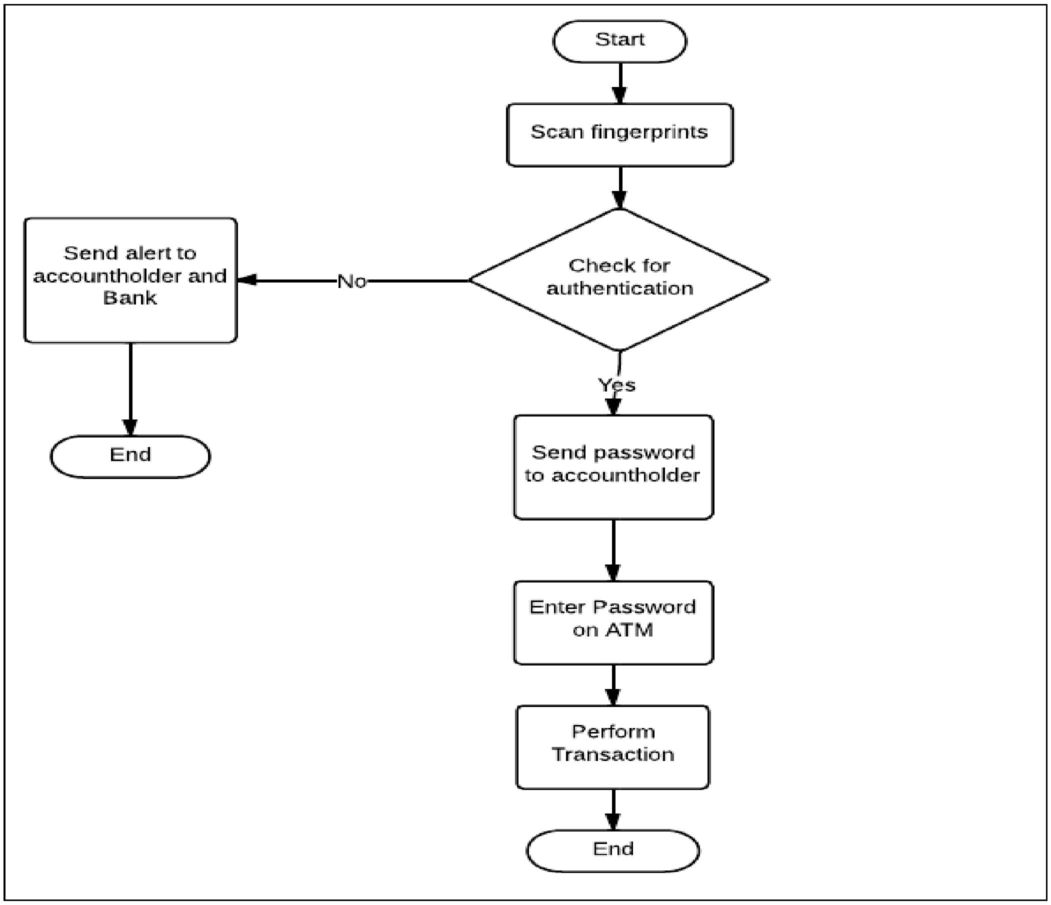


Fig. 6. Flowchart of Proposed System by Daula and Murthy (2012)

TABLE II. QUALITATIVE ANALYSIS OF PROPOSED METHODOLOGY AND OTHER APPROACHES

Methodology Approach / Analysis Criteria	**(1) Tokenless Biometric Authentication, Gelb and Decker (2011)**	**(2) Biometric Authentication using GSM, Daula and Murthy (2012)**	**(3) Proposed System**
Ease of Use	Requires just the use of fingerprints and is therefore very easily to use and reduces the burden of having to remember PINs.	Generally very easy to use and more common in system implementations. The major drawback with this system is the fact that PINs can be easily forgotten or misplaced.	Proposed system requires the use of an ATM card and the cardholder fingerprints. This system is also very easy to use. The system does not require a cardholder to remember PINs since his biometric information is used in the authentication process.
Security	Though more secured than the traditional system (2), the procedure can result in false acceptance error in cases where the number of records stored in the database is very high. This is because the identification of a user is done against a database of huge number of biometric records.	Presents a number of security concerns; ATM cards can easily be stolen and PINs are easily guessed. Additionally, some cardholders have a habit of sharing their PIN with family relatives and close friends; this reduces the security of the system.	The proposed system provides the security that comes with a biometric technology and also eliminates the likelihood of false acceptance. Hence, it only matches the cardholder's biometric information against the information stored on the ATM smart card.
Privacy	This system offers very little privacy in	This system offers more privacy to	The proposed system offers more

	terms of biometric information. The biometric information of the cardholder is stored in a database which is accessed during the point of verification.	the cardholder in terms of biometric information; this is because the system does not make use of biometrics.	privacy in terms of biometric information since the individual carries his biometric information on his ATM smart card.
Authentication Time	The tokenless system has a slow response time. This is because reading biometric information from a database and matching against each of the records being read can take some considerable amount of time.	The Traditional authentication system has a fast authentication time due to the nature of the information being transmitted to the host processor and the nature of the authentication being carried out.	The proposed system has a faster response time than the tokenless system. The system does not require access to a database and only matches the presented biometric information against the biometric information read from the ATM smart card. Furthermore, the response time compares just as equal as or better than the traditional system.
Cost	The tokenless system has a very high setup and implementation cost. There is no additional cost afterwards.	The traditional system has a low setup cost, but this cost could increase considerable when you take into account the cost of replacing missing or stolen card and the cost of generating new cards for users who have forgotten their PINs. This system could end up being a very costly approach.	Propose system has a high initial setup cost, if the cost of card maintenance is taken into consideration. This approach is more costly than the tokenless system but less costly than the traditional system; but the merits gained far outweighs cost considerations.

- Matching the biometric details an account holder against a large database of biometric details can lead to false reject errors [13].
- The system is quite slow when dealing with high capacity requirements [14].

C. Methodology Comparisons & Evaluation

The authors evaluated the proposed system in comparison to other approaches. The following criteria were used to analyze the performance of the proposed system; Ease of Use, Security, Privacy, Speed, and Cost. The authors present a tabular analysis of the methodology approaches in TABLE II. This analysis summarizes the discussions regarding methodology approaches presented in the literature, and outlines the merits of the proposed methodology over the other approaches.

VIII. Conclusion

This paper presents a design framework for the secure authentication of biometric fingerprint on ATM systems. The authors addressed the methodology that employs the use of BioHASH templates ensures an irreversible cryptographic hash function, facilitates a faster authentication, and enables an efficient framework of detecting potential duplicates of banking account holders.

The proposed framework architecture is modelled such that biometric fingerprint information of an account cardholder is captured and BioHASH tokens are generated, as a result. These tokens are then written onto the microchip of a smart (debit) card, with the biometric information discarded afterwards. At the point of ATM transaction authentication, the pre-stored BioHASH tokens on the smart card are matched against the "live" fingerprint tokens to determine legitimacy and subsequent accessibility for the supposed account holder.

The authors compared the framework methodology against other approaches and outlined the merits and suitability of their approach for delivering a robust, fast, and efficient authentication procedure on ATM systems. The following were the merits that the framework architecture discussed in this paper offers over the current systems of ATM transaction authentications; flexibility, scalability, fast user authentication, privacy preservation, and efficient security.

The analyses of the evaluation showed that the average response times of each of the procedures (and their sub-procedures) were appreciably small. These procedures were acquisition of fingerprint using scanner, completion of enrolment process, and online off-card verification. The authors' assessment indicated some areas in the design and implementation that needed improvements. For example, the adoption of a slap fingerprint scanner to increase efficiency and reduce the response time during the capturing of biometric fingerprints.

The authors' approach, thus, provides practitioners and researchers in the industry of biometric technology with methodology, procedures, and exact measures as to how successful an authentication process on an ATM system is achieved.

One critical area of future research is the implementation and testing of all the major components of this prototype design with a financial institution on a commercial scale. This will expose the design to all the practical technicalities in line with commercial use. The authors also envisage the drift of development from a standalone application system to a web application system using the Model View Controller (MVC) approach.

REFERENCES

[1] R. Silberglitt, P. S. Antón, D. R. Howell, and A. Wong, “The Global Technology Revolution 2020, In-Depth Analyses: Bio/Nano/Materials/Information Trends, Drivers, Barriers, and Social Implications,” Technical Report, [Online]. Available: http://www.rand.org/content/dam/rand/pubs/technical_reports/2006/RAND_TR303.pdf, 2006, Retrieved: 29-10-2015.

[2] D. Hutchinson and M. Warren, “Security for Internet Banking: A Framework. Logistics Information Management,” vol. 16, issue 1, pp. 64-73, 2003, ISSN: 0957-6053. DOI: http://dx.doi.org/10.1108/09576050310453750

[3] P. Magutu, M. Mwangi, R. Nyaoga, G. Ondimu, M. Kagu, K. Mutai, H. Kilonzo and P. Nthenya, “E-Commerce Products and Services in the Banking Industry: The Adoption and Usage in Commercial Banks in Kenya,” Journal of Electronic Banking Systems, vol. 2011, article ID: 678961, 19 pages, 2011, DOI: 10.5171/2011.678961.

[4] L. Bradley and K. Stewart, “A Delphi Study of Internet banking,” Marketing Intelligence & Planning, vol. 21, no. 5, pp. 272-281, 2003.

DOI: http://dx.doi.org/10.1108/02634500310490229

[5] S. Venkatraman and I. Delpachitra, “Biometrics in Banking Security: A Case Study,” Information Management & Computer Security, vol. 16, no. 4, pp. 415-430, 2008. DOI: http://dx.doi.org/10.1108/09685220810908813

[6] B. Scholnick, N. Massoud, A. Saunders, S. Carbo-Valverde and F. Rodriguez-Fernandez, “The Economics of Credit Cards, Debit Cards and ATMs: A Survey and Some New Evidence,” Journal of Banking & Finance, vol. 32, no. 8, pp. 1468-1483, 2008.

[7] S. Oko and J. Oruh, “Enhanced ATM Security System using Biometrics,” International Journal of Computer Science (IJCSI) Issues, vol. 9, issue 5, no. 3, September 2012. ISSN (Online): 1694-0814.

[8] M. O. Onyesolu and I. M. Ezeani, “ATM Security Using Fingerprint Biometric Identifier: An Investigative Study,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 3, no. 4, pp. 68-72, 2012.

[9] S. Daula and D. Murthy, “An Embedded ATM Security Design using ARM Processor with Fingerprint Recognition and GSM,” International Journal of Advanced and Innovative Research (IJAIR), vol. 1, issue 2, July 2012. ISSN: 2278-7844.

[10] S. Biswas, A. B. Roy, K. Ghosh, and N. Dey, “A Biometric Authentication Based Secured ATM Banking System,” International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 2, issue 4, April 2012. ISSN: 2277 128X.

[11] F. S. Hossain, A. Nawaz, and K. Md. Grihan, “Biometric Authentication Scheme for ATM Banking System Using Energy Efficient AES Processor,” International Journal of Information and Computer Science (IJICS), vol. 2, issue 4, May 2013. ISSN Online: 2161-5381.

[12] A. Gelb and C. Decker, Cash at Your Fingertips: Biometric Technology for Transfers in Developing and Resource-Rich Countries, Center for Global Development, Working Paper 253, 2011.

[13] G. A. von Graevenitz, “Biometric Authentication in Relation to Payment Systems and ATMs,” Datenschutz und Datensicherheit - DuD, vol. 31, issue 9, pp. 681-683, September 2007. Online ISSN: 1862-2607. DOI: 10.1007/s11623-007-0223-9.

[14] A. Jaiswal and M. Bartere, “Enhancing ATM Security Using Fingerprint and GSM Technology,” International Journal of Computing Science and Mobile Computing (IJCSM), vol. 3, issue 4, April 2014.

A New Approach for a Better Load Balancing and a Better Distribution of Resources in Cloud Computing

[1] Abdellah IDRISSI and [2] Faouzia ZEGRARI
Computer Sciences Laboratory (LRI), Computer Sciences Department,
Faculty of Sciences, University Mohammed V – Rabat

***Abstract*—Cloud computing is a new paradigm where data and services of Information Technology are provided via the Internet by using remote servers. It represents a new way of delivering computing resources allowing access to the network on demand. Cloud computing consists of several services, each of which can hold several tasks. As the problem of scheduling tasks is an NP-complete problem, the task management can be an important element in the technology of cloud computing. To optimize the performance of virtual machines hosted in cloud computing, several algorithms of scheduling tasks have been proposed. In this paper, we present an approach allowing to solve the problem optimally and to take into account the QoS constraints based on the different user requests. This technique, based on the Branch and Bound algorithm, allows to assign tasks to different virtual machines while ensuring load balance and a better distribution of resources. The experimental results show that our approach gives very promising results for an effective tasks planning.**

Keywords—Cloud Computing; Constraints QoS; Combinatorial Optimization; Task Scheduler; Exact Method

I. Introduction

Cloud computing appears as a new computer model of the company whether private, public or hybrid. It provides a paperless technical means in term of networks, servers, and storage. The cloud computing is developed primarily through distributed computing, parallel computing and grid computing. Distributed computing allows the cloud computing, to decompose a global operation into several tasks, and then send them to processing systems. The needs of the internet users are often various and depend on the tasks. However, resource planning becomes more complex in an environment composed of heterogeneous resources, and it depends on the requirements of users. Cloud computing should then integrate the resources of heterogeneous networks to minimize the completion time of all tasks and maximize resource utilization [1]. As the Quality of service (QoS) represents standards of satisfaction of using Cloud services, it is a question to coordinate the different resources and to optimize their planning.

Cloud computing is an offer for all the ICT requirements. Defined by the NIST (National Institute of Standards and Technology) [3] as a service on demand, the cloud provides access to a shared pool of configurable computing resources. On the economic front, this new model has an important budgetary and financial impact. It is a strategy to reduce costs and maximize the return on investment. Therefore, provide more flexibility and agility will be more efficient for economy. The rest of the paper is organized as follows: the next section describes the concept of cloud computing, Section 3 presents the QoS models and its various aspects. Then, we explain principles of the Genetic Algorithm in section 4, and those of the Branch and Bound followed by our propositions in section 4. In section 5, we present experimentations and results and finally, we conclude in section 6.

II. Concept of Cloud Computing

A. *Concept of Cloud Computing*

Cloud computing is not entirely a new approach. Computing has emerged in the 60s and was marked by "computing on demand", proposed by John McCarthy which he predicted in a speech that computing will one day be a public utility [4, 5]. The 80's were stacked by concepts of virtualization. Moreover, a few years ago, the ASP model (Application Service Provider) was used to propose an application as a service [5]. Through techniques of virtualization and the grid architecture [6] which allows the rise to power of a service, the idea of john is now concretized by the emergence of cloud computing. In that way, Cloud Computing was allowed to pass from the fixed price approach to the pay-as-you-go mode, a choice payment model in the disposition of the informatics demands [7, 8].

To address the problem of performance degradation of shared resources, the cloud systems use dynamic scheduling of virtual machines with the technique of dynamic migration. However, virtualization makes possible the rapid replacement of a server in a cloud without charge or major damage. It dynamically responds when allocating resources is needed, and when adapting applications with computing resources, storage and network. This option distributes workload according to the requirements requested. The Cloud belongs to the virtualization technologies and therefore provides infrastructure and platforms on demand.

This cloud model is defined by five fundamental principles, three service models and four deployment models.

The fundamental principles are the basis of all cloud computing architecture: shared resources, elasticity, self-service, payment for the use and accelerating the speed [5].

B. *Technical models [7, 10]*

The structure of the cloud consists of three layers: Infrastructure, Platform and Software.

- IaaS (Infrastructure as a Service) is the lowest layer of cloud computing. It provides all the hardware equipment that the company can rent in remote datacenters, for the need of its applications to run the IT.

- PaaS (Platform as a Service) is the service, which provides a development environment online. PaaS is the platform of execution, deployment and development of applications.
- SaaS (Software as a Service) is the final layer of cloud. It provides applications provided on request. It is a service ready to be consumed right away, accessible via Internet.

C. *Deployment models*

The main deployment models as presented in [4, 10] are:

- Public cloud or external cloud: it is a structure managed by a provider. Services are provided to various organizations via the internet ;
- Private cloud or internal cloud: the infrastructure of this type of cloud is rented by a company and is only operated by operational units via its intranet. Its infrastructure is not mutualized ;
- Hybrid Cloud: it's a mixed structure. It combines the internal resources of private cloud with external resources of public cloud ;
- Community Cloud is implemented by specific community organizations sharing common interests.

D. *Advantages and disadvantages*

The benefits of cloud computing are many through decentralization of storage space and the pooling of IT resources. We cite some advantages as: Adaptability of resources as needed [9, 10, 11, 18, 19, 20], Applications automatically benefit from security improvements and performance, Ensure high availability of services and data, and reduce risks [12]. The major disadvantage of cloud computing is related to its security. It must maintain security of information stored in the clouds, in terms of integrity, risk of intrusion, control of the documents on their storage and geographical location

III. MODELS QOS (QUALITY OF SERVICE)

The cloud is the lever of development to meet the needs of clients. It guarantees perfect quality of service through load balancing across several servers or data centers, and again through the implementation of procedures to restore applications and data backup in case of disaster. In effect, with the multiplication of services in the cloud, several questions can be asked about the quality of service (QoS) rendered [18, 19, 20]. The QoS refers to the ability to respond with quality to user needs and to provide a service according to the requirements in terms of response time, bandwidth, availability, etc. So, the differentiator between cloud computing offerings will be the quality of service provided. The grid concept allows the growing computing power, which need monitoring the quality of provided computing resources [13]. Scheduling of resources takes account of QoS constraints on both aspects: at the user level and the system level [14].

The scheduling problem can be modeled by these three criteria, constituting a multi-objective function defined by the weighted summation of the execution time, cost and load, as defined in [8].

$$\mathrm{M(x)} = (\omega_1 \times \mathrm{Time}) + (\omega_2 \times \mathrm{Cost}) + (\omega_3 \times \mathrm{Load}) \qquad (1)$$

With : $\omega_1 + \omega_2 + \omega_3 = 1$
ω_i : weight coefficient of each indicator

The Time Indicator refers to the execution time, it comprising the processor capacity (r_cpu) and bandwidth (r_comm). We will use the same formulas expressed by the authors in [8].

$$\mathrm{Time} = (Time_{exec}) + (Time_{comm}) \qquad (2)$$

$$Time_{exec} = \mathrm{rq}_{instruction_{\mathrm{count}}}/\mathrm{r}_{\mathrm{cpu}} \qquad (3)$$

$$Time_{comm} = \mathrm{rq}_{\mathrm{size}}/\mathrm{r}_{\mathrm{comm}} \qquad (4)$$

The rq_instruction_count is the length of the task, and the rq_size is the size of the data file.

In calculating costs, we refer to the four aspects of cost including CPU, memory, disk and bandwidth. We can define a billing formula as follows: the CPU cost versus time.

The cost is expressed by the following equation:

$$\mathrm{Cost} = Cost_{CPU} + Cost_{Ram} + Cost_{BW} + Cost_{Stor} \qquad (5)$$

The load indicator includes three parameters, which are respectively the CPU usage (Load_cpu), the memory usage (Load_mem) and the use of the bandwidth rates (Load_br) [8].

We seek to minimize the maximum load, which corresponds to the load balancing problem.

The weighted function, which allows balancing between the utilization of CPU, memory and bandwidth is given hereafter.

$$\mathrm{Load} = 1 - \prod_{\mathrm{k}=1}^{3} (1 - \mathrm{Load_k})^{\omega_{\mathrm{LK}}} \qquad (6)$$

The variables Load_cpu, load_mem and Load_br are determined as defined in [8, 15].

Through the technology of virtualization, each computing node is defined by a set of attributes constituting the Resource Information (RI) from which the task will choose for scheduling, comprising the CPU calculation ability, the memory size, the price and the load capacity of the resource. It is expressed in [8] as follows:

$$\mathrm{RI} = \{\mathrm{r_{cpu}}, \mathrm{r_{mem}}, \mathrm{r_{stor}}, \mathrm{r_{comm}}, \mathrm{r_{cpu_{cost}}}, \mathrm{r_{mem_{cost}}}, \mathrm{r_{stor_{cost}}}, \ \} \qquad (7)$$

The information of the task is composed of attributes allowing demand for resources to accomplish a task [8].

$$\mathrm{rq} = \{\mathrm{rq}_{\mathrm{cpu}}, \mathrm{rq}_{\mathrm{mem}}, \mathrm{rq}_{\mathrm{stor}}, \mathrm{rq}_{\mathrm{comm}}, \mathrm{rq}_{\mathrm{instruction}_{\mathrm{count}}}, \mathrm{rq}_{\mathrm{size}}\} \quad (8)$$

- r_cpu: calculation ability whose unit is MIPS;
- r_mem: indicates memory size provided by the node whose unit is MB;
- r_stor: means storage space of data provided by the node. Its unit is GB;
- r_comm: refers to capacity of data transfer that node can provide. Its unit is MB/S;
- r_cpu_cost: indicates the price of the processor ;
- r_mem_cost: indicates the price of the memory. It consider 1024 MB as calculation reference;
- r_stor_cost: indicates the price of data storage. It consider 100 GB as calculation reference;
- r_comm_cost: indicates the price of bandwidth. Its unit is 1MB/S.

Based on the multiple QoS constraints environment, task scheduling of cloud computing is to allocate tasks on the appropriate resources. We focus our research fields on the task scheduling, which is one of combinatorial optimization problems. The goal is to order the execution of operations on different virtual machines VMs in the Cloud Computing environment, so as, to minimize the execution time and cost while ensuring load balancing, as described in [8, 14]. In our case, we propose to use a Branch-and-Bound algorithm and we will compare the results with genetic algorithm. We recall hereafter the description of Genetic Algorithm and Branch-and-Bound.

IV. Genetic Algorithm in Cloud Computing

A. Problem

In cloud computing, users submit different kind of tasks whose requirements can be defined according to the corresponding weights to the execution time, cost and load. Given the diversity of virtual machines hosted in the cloud makes it difficult to allocate these tasks to the appropriate machines. So, to deal with this problem, we choose some computing nodes that meet the requirements in order to form the initial population of our algorithm. Then, we run the simulation that we have programmed in Java in order to obtain a satisfactory allocation of system resources.

B. Analysis of the problem

The principle of genetic algorithm is to evolve an initial population of individuals, by successive generation, based on the mechanism of genetic operators: crossover, mutation and selection.

We start by creating a list of virtual machines: vmList and a list of tasks: taskList. We propose that vmList is the initial solution of the problem, (SolutionList=vmList), and will contain, for each iteration, the most suitable generation.

This allows us to choose among the individuals, those that can reproduce and can undergo at the crossing. Among the various methods of selection, we opted for the random selection. The probability of selection of each individual is 1/PopSize where PopSize represents the size of the population.

The crossing allows generating one or two children by an exchange of information between two parents. As for mutation, it aims to modify a random part of the population, causing a perturbation of the gene of the chromosome, with a low rate in order to avoid a random dispersion of the population.

C. Genetic algorithm

```
Begin
Function algoGenetic(MaxGeneration : int, vmList :
List<Vm>, taskList : List<Task>,  allocList :
List<Allocation>)

nbGeneration : int : declaration of a generation counter
bestAlloc ← null : intialize best chromosome
constPopulationInitiale(vmList, taskList, allocList) :
creating the initial population
tabFits : Fitness[] : initialize table tabFits to store fitness of
chromosomes

While(nbGeneration < MaxGeneration)

 allMi : List<Double> : initialize list to store objective
 function of each chromosome


 for I ← 0 to allocListSize then              }  calculate fitness
     tabFits ← fitness(allocList(i)) ;        }     of each
     allMi ← tabFits[i].getMi());             }    allocation
 end for
 SortMi(allMi) : sort list allMi              }  select a couple
 ch1 ← selection(allMi,allocList)             }  of parents: ch1
 ch2 ← selection(allMi,allocList)             }     and ch2

 parent1 ← RandListCh1                        }  distribution of
 parent2 ← RandListCh2                        }   each parents


 childrenAlloc ← Crossing(parent1,parent2) : apply
                crossover on two parents to create the children
 childFit=new Fitness(childrenAlloc) : calculate fitness of
                                       children

 if(childrenAlloc.getMi () <= listAveragesSort.getLast())
     replace bad parent by child in the list allocList
 EndIf


 childrenMutate ←  mutation(childrenAlloc) : apply
                  mutation on created child according
                  mutation probability
 SortMi(allMi) : sort list allMi
 Take the index for first element of the list allMi

compareAllocation(firstChildIndex , bestAlloc) : Compare
                          the best allocation, with the allocation
                already found
   Replace the older generation by new generation
   MaxGeneration++ : increment the number of generation

  EndWhile
 EndFunction
End
```

V. Branch and Bound in Cloud Computing

In general, the exact methods are based on finding minimum cost solutions to solve NP-hard problems [16]. These methods are implicit enumeration techniques based on the branch and bound method. They allow exploring all branches intelligently by pruning subassemblies, which do not lead to good solutions [17].

Two bounds define this technique: upper and lower. At each vertex is associated a reduction function of the cost computed. The optimal solution with respect to solutions already found is the solution of the initial problem.

A. *Presentation*

Combined with QoS constraints and the concept of optimization, we use the techniques of Branch and Bound for the scheduling tasks problem in the cloud computing environment in order to obtain a distribution scheme satisfying resources. This approach aims to allocate the tasks on the virtual machines which are more appropriate to the requirements expressed by users. It provides better results in terms of performances and costs.

B. *Analysis of the problem*

Generally, formalizing a combinatorial optimization problem consists in incorporating to the problem an objective function, the constraints of the problem and assigning values to variables, which must be defined to determine the set of solutions respecting the constraints.

The function M(x), as given in formula (1), is composed of three variables namely: time, cost and load. It comes to a problem that integrates multiple criteria. These criteria, being often contradictory, can render resolving the problem more difficult.

Then, we have recourse to a multi-objective optimization process, which consists of simultaneously optimize multiple objective functions. Our work proposes to adapt the branch and bound algorithm for solving this sort of problem. Within the implementation of this algorithm we propose to decompose this function M(x) into three sub-objective functions: **$fobj_{Time}$**, **$fobj_{Cost}$** and **$fobj_{Load}$**.

We assume execute n tasks in m virtual machines. The algebraic formulation of the problem is as follows:

- P_{ij} execution time of task j in the virtual machine i ;
- C_{ij} cost of using resources;
- L_{ij} load in the resource i running task j

Thus, a permutation matrix "X" is defined to ensure the assignment of a task to a single processor, such as:

$$\begin{cases} x_{ij} = 1, & \textit{if task } j \textit{ is executed on} \\ & \textit{virtual machine } i \\ x_{ij} = 0, & \textit{otherwise} \end{cases} \qquad (9)$$

We seek to:

- minimize the execution time, which is defined as the greatest completion time of all tasks in all machines. From the formula (6) described in [8], we have :

$$\boldsymbol{fobj_{Time}} = max_{1\le i\le m} \sum_{j=1}^{n} P_{ij} * x_{ij} \qquad (10)$$

VARIABLES : $V_{Time} = \{r\_cpu, r\_comm\}$

DOMAINS:

r_cpu : $[20000, 50000]$ MIPS
$Lenght$: $[100, 10000]$ MI
$File\ Size$: $[35, 300]$ Ko
$Output\ Size$: $[35, 300]$ Ko
Bw : $[0, 10]$ MB/S

CONSTRAINT:

We can assume a maximum time not to be exceeded

- minimize the cost of using resources, which is defined as the summation cost of all tasks in all machines. the formulas are as following :

$$\boldsymbol{fobj_{Cost}} \sum_{1\le\, i\, \le m} \sum_{1\le\, j\, \le n} C(i,j) * x_{ij} \qquad (11)$$

VARIABLES :

$$V_{Cost} = \begin{Bmatrix} r_{cpu_{cost}}, r_{mem_{cost}}, r_{comm_{cost}}, rq_{cpu}, rq_{mem}, \\ rq\_comm \end{Bmatrix}$$

CONSTRAINTS:

Cost <= Budget set by the client

We can assume other constraints.

- minimize system load, which is defined as the largest load in all machines.

$$fobj_{Load} = max_{1\le i\le m} \sum_{j=1}^{n} L_{ij} * x_{ij} \qquad (12)$$

With:

$$Load = 1 - \left[\left(1 - Load_{cpu}\right)^{\omega L1} (1 - Load_{mem})^{\omega L2} (1 - Load_{bw})^{\omega L3} \right] \quad (13)$$

VARIABLES : $V_{Load} = \{Load\_cpu, Load\_mem, Load\_br\}$

C. Branch And Bound algorithm

```
Function AlgoBnB
  Begin
  vmList : List<Vm>
  taskList : List<Task>
  BestSol ← Null : Initialize a list that will contain the
                   optimal solution
  constPopulationInitiale(vmList, taskList) : create initial
  population
  vm : VM : Virtual Machine v

     ubTime ← ∞     }
     ubcost  ← ∞    }  initialize Upperbound
     ubLoad ← ∞     }  to a higher value
     lbTime  ← null
     lbcost   ← null }
     lbLoad  ← null  }  initialize lowerbound
                        to zéro

   Function calcrecursif(vmList : List<Vm>, currentbest :
   List<Vm>): this method selects (for each level) a node
           having the smallest evaluation as vertex to explore

  vmlistlocal ← vmList.clone()
  cureentbestlocal ← currentbest.clone()

  IF(vmListlocal.size()==1)
       calctrier(vmListlocal, v) : Calculate lower bound of
       the machine and check if the terminal is improved:

       IF(lbTime<ubTime && lbcost<ubCost &&
                 lbLoad<ubLoad) then

            ubTime ← lbTime           }  Updating optimal
            ubcost← lbCost            }  solution and its
            ubLoad ← lbLoad           }  terminals
            BestSol ← currentbestlocal   upperbound
       EndIF
  EndIF
  Else                                }  Calculate lower
       calctrier(vmListlocal, vm)     }  bound of each vertex
                                         and check if there
                                         are improvement

   test if evaluation of the node is better than the current
   evaluation

       IF(lbTime<ubTime && lbcost<ubCost &&
          lbLoad<ubLoad) then
               currentbestlocal ← vm : Add the partial
               solution in currentbestlocal
               subVms ← genererSublist() : generate new
               subset which does not contain the processed
               node
               calcrecursif(subVms, currentbestlocal) :
       EndIF
       Else
          prune the node; return
   End
EndFunction
```

VI. Experimentation and Results

We implemented in Java our approach based on the branch-and-bound applied to the task scheduling, in cloud computing and considering QoS constraints. We compared the experimentation results with those of the genetic algorithm. To do so, we randomly generated a number of 4 virtual machines and a number of tasks from a range of values {10, 100}. Table 1 and Table 2 show the configuration of the resources. We put the weighting coefficients with values between 0 and 1, corresponding to the user's cpu rate, memory, and bandwidth according to our needs.

For the calculation of resource utilization costs, we proposed a billing formula per unit. For each second of CPU time, and for each 1024 MB occupied, and for every 100 GB of data space, and for each 1MB/s taken of the bandwidth; will be billed just one unit cost.

TABLE I. Host Configuration

N° of Host	1
Processing Power (MIPS)	150 000
RAM (MB)	256 000
Bw (Mb/s)	2000

TABLE II. Virtual Machines Configuration

Virtual machines	VM1	VM2	VM3	VM4
P.Power (MIPS)	1024	4096	4096	4096
RAM (MB)	4000	3000	5000	5000

The results of this experimentation are shown in Figure 1, Figure 2 and Figure 3.

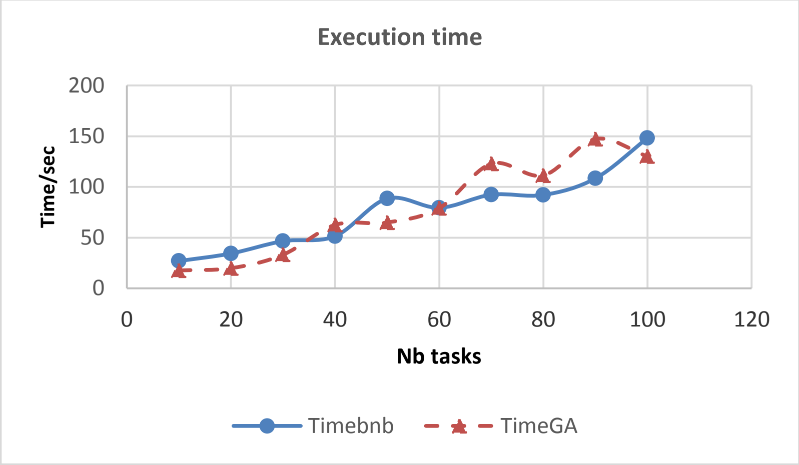


Fig. 1. Execution time graphics

Figure 1 illustrating the evolution of time according to the number of tasks, shows that branch and bound algorithm (BnB) gives better results. The graphic increases gradually. For a number of tasks lower than 30, the genetic algorithm is slightly better. The intersection of the two graphics appears from this point, where time BnB has a tendency to decrease. Our model gives a good prediction; and becomes better beyond 30 tasks than the genetic algorithm.

The Figure 2 for cost shows that the curve is nearly similar to the Figure 1. The curves of time intersect at 35 tasks. From this point, the cost of BnB is less than that of the genetic algorithm. This proves that BnB can ensure optimal solutions and better satisfy cloud computing users.

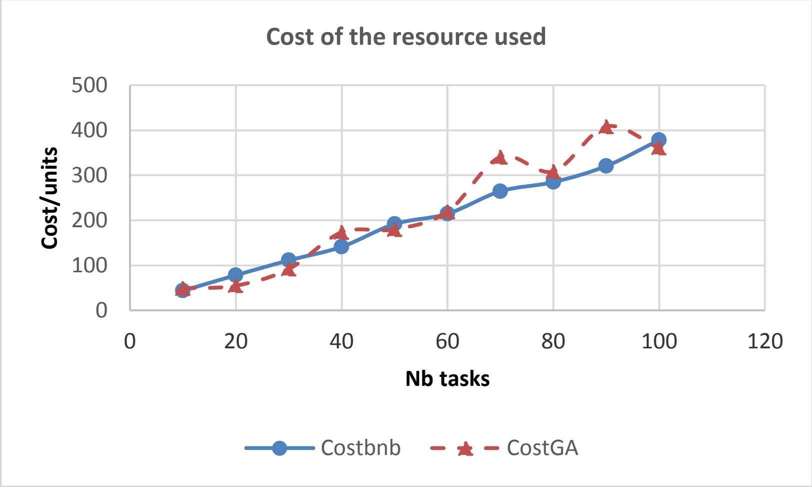


Fig. 2. Cost graphics

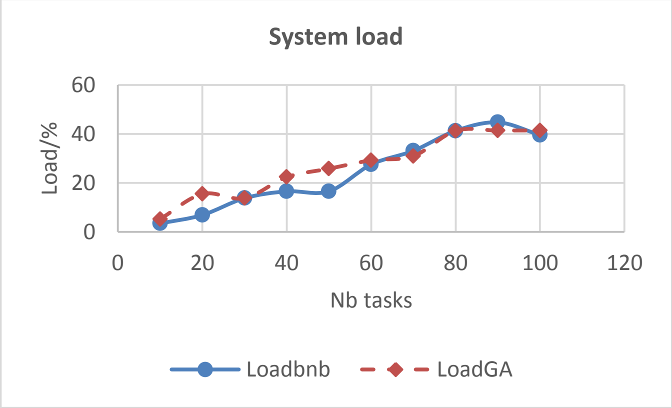


Fig. 3. Load graphics

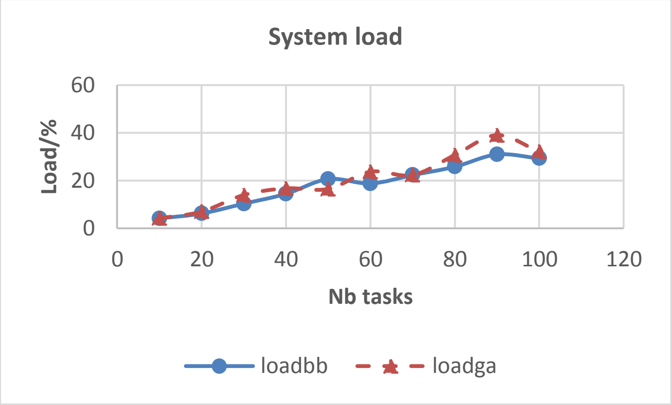


Fig. 4. Load graphics

The system load is a constraint in the cloud. It can unbalance the resource scheduling.We can deduce from Figure 3, that the system load with BnB is improved. The load increases more and more when increasing the number of tasks, the graphics show that the BnB consumes less load compared to the genetic algorithm. Other tests were performed. Among the obtained results, the graph of Figure 4, confirms that the BnB algorithm remains better and consumes less load though we increase the number of tasks.

Therefore, we can conclude that both genetic and BnB algorithm can solve optimization problems effectively and better meet the requirements of quality of service. Thus, when it is a small population of tasks, we can opt for the BnB algorithm, and for a population from a hundred tasks, the genetic model is more suitable and outperforms the BnB algorithm approach. Finally, the virtual machine is selected based on the cost and processing power.

From these comparisons, we can conclude that the Branch and Bound algorithm can solve optimization problems effectively and better meet the requirements of quality of service. Thus, when there is a small population of nodes, we can choose the BnB algorithm, and for a large population, the genetic model is more suitable and surpasses the BnB approach. Finally, all these results allow us to determine the capacity of the branch and bound algorithm, which is limited to a number of nodes not exceeding 12. Beyond that, the running time increases rapidly and can cause program shutdown.

VII. Conclusion

In this paper, we studied an algorithm for task scheduling in a cloud computing environment using an exact approach. Our approach is based on the branch-and-bound algorithm, incorporating the QoS constraints on both aspects: the user aspect and the aspect of system load balance. We followed a rational approach, based on a comparative study with the focus on the value analysis. We showed the interest of this algorithm through experimental results that allowed us to evaluate the performance of cost and system load. In comparison with the genetic algorithm, the Branch-and-Bound technique gave better results for a small population of nodes in terms of time, cost and load balancing. We can conclude that the establishment of an effective task scheduling system can meet the requirements of users, with a good use of resources, and improving the overall performance of the cloud computing environment. Thus, scheduled tasks are regarded as a management tool for cloud servers.

References

[1] Sandeep Tayal. "Tasks scheduling optimization for the Cloud Computing Systems". International Journal of Advanced Engineering Sciences and Technologies (IJAEST). Vol. 5, Issue No. 2, 111 – 115, pp:1-15, 2011 http://www.ijaest.iserp.org.

[2] Fei Teng. Thèse. "Management des données et ordonnancement des tâches Sur Architectures Distribuées ". Octobre 2011

[3] Vincy Goyal and Ruchi Dave. "A Survey on Cloud Computing Services". International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 1, Issue 4, November – December 2012.

[4] I. Foster, Y. Zhao, I. Raicu and S. Lu. "Cloud Computing and Grid Computing 360-Degree Compared,.IEEE Grid Computing Environments Workshop, IEEE Press, 2008

[5] Wygwam Tm. "Cloud Computing:Réelle révolution ou simple évolution". 2011-Bureau d'expertise technologique.

[6] I. Foster. "There's Grid in them thar Clouds. January 2008". http://ianfoster.typepad.com/blog/2008/01/there s-grid-in.html

[7] Srinivasa Rao, Nageswara Rao, Kusuma Kumari. "Cloud Computing : an overview". Journal of Theoretical and Applied Information Technology (JATIT), 2009.

[8] Guang Liu, Chen Yang and Daoguoli. "Scheduling research based on genetic algorithm and Qos constraints of Cloud Computing". Journal of Theoretical and Applied Information Technology, 10th May 2013. Vol. 51 No.1, pp. 92–95.

[9] Abdellah Idrissi and Manar Abourezq. "Skyline In Cloud Computing". Journal of Theoretical and Applied Information Technology, Vol. 60, No. 3, February 2014.

[10] Pradip Patil, Gurudatt Kulkarni and Amruta Dongare. "Cloud Computing an overview". International Journal of Modern Engineering Research (IJMER). Vol.2, Issue.2, Mar-Apr 2012 pp-380-382.

[11] Cisco, "Les bases du cloud computing : revaloriser les technologies de l'information" © 2011 Cisco Systems, Inc. et/ou ses filiales. 2mai 2011

[12] Emmanuel Boucher. "Software as a Service". Avril 2009.

[13] Kadda Beghdad ,Bey, F. Benhammadi and Faouzi Sebbak. "Fuzzy Subtractive Clustering Based Prediction Approach for CPU Load Availability". Laboratoire de Systèmes Informatiques, Ecole Militaire Polytechnique, Alger, Algérie. CLOUD COMPUTING 2013: The Fourth International Conference on Cloud Computing, Grids, and Virtualization.

[14] Yogita Chawla and Mansi Bhonsle. "Dynamically optimized cost based task scheduling in Cloud Computing". International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 2, Issue 3, May – June 2013.

[15] Thomas Heinze, Yuanzhen JI, Yinying Pan, Franz Josef Gruznzberger, Zbigniew Jerzak and Christof Fetzer. "Elastic Complex Event Processing under Varying Query Load" BD3 @ VLDB, pp. 25-30, 2013.

[16] Jens Clausen , "Branch and Bound Algorithms, Principles and Examples", Department of Computer Science, University of Copenhagen, Universite tsparken 1, DK-2100 Copenhagen, Denmark., March 12, 1999.

[17] Brandon Malone, Changhe Yuan, Eric A. Hansen and Susan Bridges, "Improving the Scalability of Optimal Bayesian Network Learning with External-Memory Frontier Breadth-First Branch and Bound Search". Department of Computer Science and Engineering - Mississippi State University - Mississippi State, MS 39762 – 2011.

[18] Manar Abourezq and Abdellah Idrissi. "Integration of QoS Aspects in the Cloud Service Research and Selection System". International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 6, Issue 6, June 2015.

[19] Manar Abourezq and Abdellah Idrissi, "A Cloud Services Research and Selection System", International Conference on Multimedia Computing and Systems (ICMCS), 2014

[20] Manar Abourezq and Abdellah Idrissi, "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

Distance and Speed Measurements using FPGA and ASIC on a high data rate system

Abdul Rehman Buzdar, Liguo Sun, Azhar Latif, Abdullah Buzdar
Department of Electronic Engineering and Information Science
University of Science and Technology of China (USTC)
Hefei, Peoples Republic of China

***Abstract*—This paper deals with the implementation of FPGA and ASIC designs to calculate the distance and speed of a moving remote object using laser source and echo pulses reflected from that remote object. The project proceeded in three phases for the FPGA implementation: All-in-C design using Xilinx Microblaze soft core processor system, an accelerated design with custom co-processor and Microblaze soft core processor system, and full custom hardware design implemented using VHDL on Xilinx FPGA. Later the complete system was implemented on ASIC. The ASIC implementation optimized the modules for area and timing for a 130nm process technology.**

Keywords*—*Distance; Speed; FPGA; MicroBaze; Co-Design; ASIC

I. Introduction

Distance and speed measurement systems are widely used in many areas including automobiles, defense etc [1-15]. The system can measure the time interval between two laser pulses, one reference pulse which is sent out by the system and one echo pulse which is reflected back to the system. With that time information the system should calculate the distance to the object on which the echo pulse is reflected. The system level view of the project is shown in Fig. 1 and the reference and echo signals are depicted in Fig. 2. There are to be two phases in the project, in the first phase the target hardware is a Xilinx FPGA [16], in the second phase the target hardware is an application specific integrated circuit (ASIC) solution where VHDL and standard cells is to be used. The target process is a 130nm CMOS process from the foundry ST Microelectronics [20]. Cadence [21] EDA tools are used for the ASIC implementation. The FPGA system implementation phase can be further divided into three sub phases. The first of these phases aims to deliver a software only product, where the entire system is written in C programming for Xilinx MicroBlaze soft processor system [17]. The second is a mixed hardware-software implementation where a part of the system is implemented in C and part of it in hardware. The hardware is to be designed using VHDL hardware description language. The third and final product is to be implemented completely in hardware and implemented on Xilinx XUP Virtex-II Pro Development System [18]. Lacking a real laser detector, a laser emulator is to send different test vectors to the distance measurement system. It should feed the digital filter block through two channels, each channel operating at 100 MHz. A test vector generator emulating the laser detector i.e emulate the laser signal pulses “Reference” and “Echo” is developed in VHDL and is named “AD model”. Each measurement should comprise of 256 samples. Finally as the system should be able to measure speed and the emulator should also be able emulate movement. The second block is the digital filter, it is a five tap correlation filter. The signals must be filtered using a digital filter shown in Equation (1) to suppress noise. where A-E are the filter coefficients. Initial filter coefficients have been provided but the user should be able to change coefficients during use.

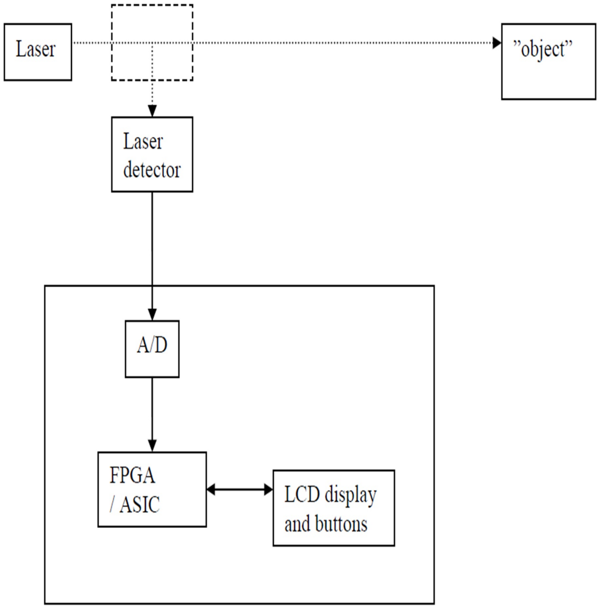


Figure 1: System Level View of Project

After filtering, the maximum point of the reference signal is found by searching for the maximum value of $d_1[i]$ in the range 0 - 20 samples (i = 0 .. 19). 20 samples are supposed to be the longest possible delay between “laser-trig” goes high and the laser pulse is transmitted. In the same way a possible echo signal is searched in the interval 21 - 255 samples (i = 21 .. 255).

$$\begin{aligned} d_1[i] = A * d_0[i-2] + B * d_0[i-1] + C * d_0[i] \\ + D * d_0[i+1] + E * d_0[i+2] \quad (1) \end{aligned}$$

To increase the resolution of i an interpolation using the two values surrounding the maximum value $d_1[i]$ is performed according to Equations (2), (3), and (4). This gives the final time sample point j.

$$b = d_1[i-1] - d_1[i+1] \tag{2}$$

$$c = 2*(d_1[i-1] - 2*d_1[i] + d_1[i+1]) \tag{3}$$

$$j = i + b/c \tag{4}$$

The third block is the distance and speed measurement. Main requirement for this block is that the distance should have a precision of one decimal. The distance to the object can be found by calculating the time difference between the reference and the echo signals. Assuming the sampling frequency is 200 MHz and using the speed of light, we get Equation (5):

$$Distance = (j_{echo} - j_{ref}) * c/(2*200*10^6) \tag{5}$$

Fourth block is a user interface utilizing an LCD for printouts and a keypad for input. Maximum range for the system should be 250 meters. Measurements should be possible at a 100 KHz repetition frequency, that is each measurement should be performed in less than 10us.

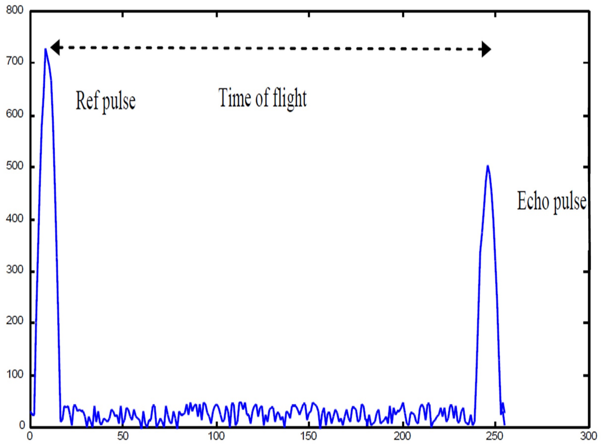


Figure 2: Reference and Echo Signals

II. Software Implementation

In this phase everything was implemented in C programming language. The C code was first tested on personnel computer (PC) to see how the filter and other blocks are working. This implementation made the idea behind the project clear. After successfully testing it on PC, the same approach was used for Xilinx MicroBlaze soft processor system [17]. As floating point operations are expensive therefore it was decided not to use floating point unit. Due to the lack of a floating point unit a fixed point system was implemented. The actual speed of light is 299792458 but it was rounded off to $3*10^{-8}$. This rounding off resulted in a little deviation in the decimal part and that was expected. We first printed the results using Hyper-terminal by connecting serial port of Microblaze with the serial port of PC. Later we integrated LCD to display the results. Fig. 3 shows the block level diagram of software implantation.

In this phase LCD was derived and values were received from the keypad using Software routines. Timers were also used for profiling. Fast Simplex Link (FSL) [19] was used to get values as test vectors. FSL is 32 bits wide, but only ten bits were used as the test vectors are 10 bits wide. In this phase everything was sequential. The Fig. 4 shows the software implementation flow. The sample test vectors were buffered first in an array which was then fed to the filter. After the filtration the max value was found and index was located, later interpolation was carried out and finally distance was calculated. The profiling was done to figure out the time consuming parts of the code. The table I shows the results from profiling.

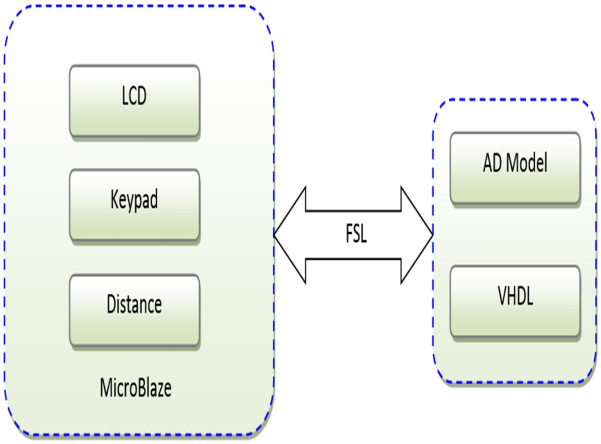


Figure 3: Block diagram of software implementation

Table I: Result of C routine Profiling

C routine	Time[us]
Read sample vector	25
Filter	789
Find Ref	5
Find Echo	14

A. Keypad Implementation

The keypad has four rows and four columns. The columns were sourced and output was sensed on rows. The keys acts like simple switches. The signal put on the columns was active low and then the rows were sensed. If there is a low signal present on the any of the row then the routine figures out which key is pressed. A routine was included for resolving debounncing issue. As the switches are kind of mechanical switches so whenever any of the key is pressed there is kind of fluctuation at the output pin and the routine can register the key multiple times.

To fix this de-bouncing problem this routine was included. The routine makes sure that it should register the key once until the key is depressed. Four keys were used for changing test vector they are 1 2 3 A as shown in Fig. 5. The Fig. 6 shows the flow for Keypad routine.

B. Distance Calculation

For the distance calculation FSL link was used to get the vector values from AD model, then they were fed to the distance calculation routine. After getting these vectors, filtering and interpolation was carried out. There was little variation in decimal place as the speed of light was rounded off. Table

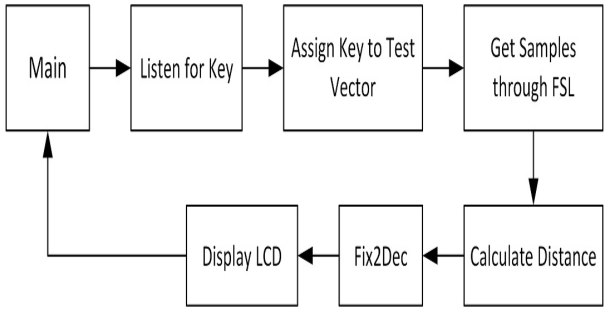


Figure 4: Software Implementation Flow Chart

Figure 5: Keypad Keys

I shows the result of profiling on the Distance calculation routine. The Fig. 7 shows flow for Distance calculation routine.

1) Fix2Dec: This routine takes the digit from the distance calculation routine and separates it into integer and decimal part. These values are then sent to the LCD routine to be displayed on LCD.

C. LCD Implementation

The LCD used is a 16x1 display format. It contains three control signals and 8 bit wide data bus to receive commands and data. The control signals are Enable, Read/Write and Register Select. Data and commands should be latched on the Enable signal. For displaying to the LCD, write signal should be low and for command and data Write Register should be Low and High respectively. Fig. 8 shows the flow for LCD routine.

III. Mixed HW-SW Implementation

In mixed HW-SW Implementation Phase some parts of the project were implemented in software and some in Hardware. The Hardware-Software co-design is a well established technique, which improves the performance of the system [22-36] . The main interface i.e. Menu System, LCD and keypad was implemented using Software routines and the distance chain was implemented in hardware as shown in the Fig. 11. Here is a brief description of each routine.

A. Main Interface Implementation

The main interface involves a menu system which takes user input through a keypad and displays the results on the LCD. The menu system is implemented in software. The menu has the options to change coefficients, test vectors, speed and display the calculated value of distance on the LCD. Fig. 9 shows the Main Interface flow chart. The same code was used to implement the Keypad and LCD as described in software implementation phase.

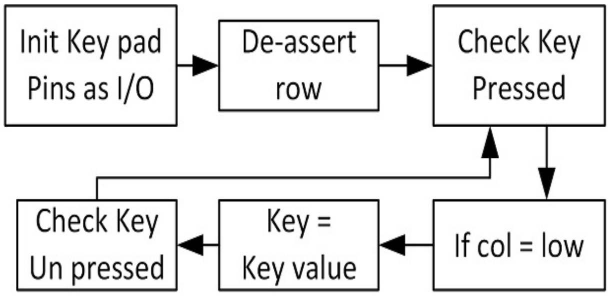


Figure 6: Keypad Routine Flow Chart

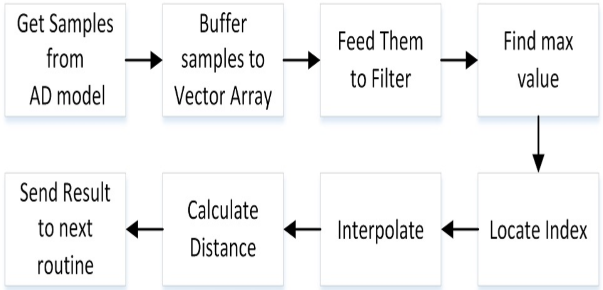

Get Samples from AD model
Buffer samples to Vector Array
Feed Them to Filter
Find max value
Locate Index
Interpolate
Calculate Distance
Send Result to next routine

Figure 7: Distance Chain Routine Flow Chart

B. Menu Routine Implementation

In mixed HW-SW Implementation Phase the main interface is implemented using only four buttons i.e. plus, minus, enter and back. The reason of using less number of keys is to minimize the number of pins out from our design.

When the device is turned ON the user can choose between different options from the main menu and they are displayed on the LCD i.e. change test vectors, coefficients and speed. To move backward and forward in the menu the plus/minus keys are provided to the user and to select any of these options enter key is used.

C. Distance Chain Implementation

In mixed HW-SW Implementation Phase we decided to implement the whole distance chain in VHDL and only keep non-timing sensitive stuff inside the MicroBlaze. The reason for doing this is that we know that we would have to do it in complete hardware Implementation Phase and this way could save time although it might be harder. The Fig. 10 shows the overview of the datapath of distance chain for mixed hw-sw implementation. The Distance chain is compromised of these blocks:

D. AD Model

Very small changes was done to this block from software implementation phase, we implemented a start signal and then a start filter signal to show when the filter would have its first two values. Together with the start signal the test vector chosen is sent.

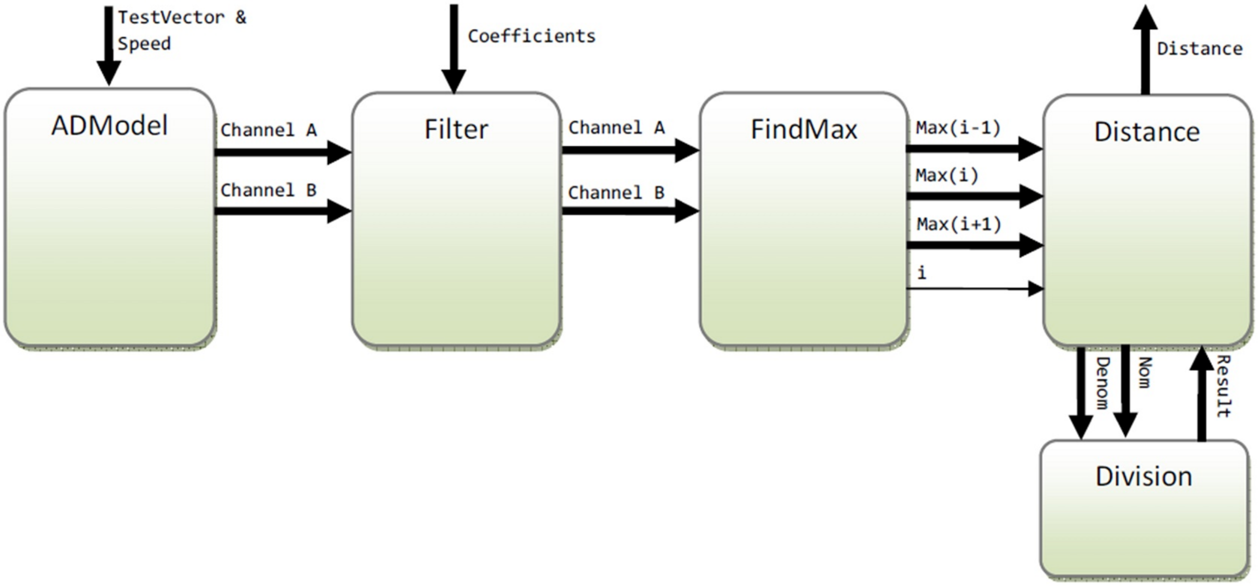


Figure 10: Overview of the datapath of distance chain for mixed hw-sw implementation.

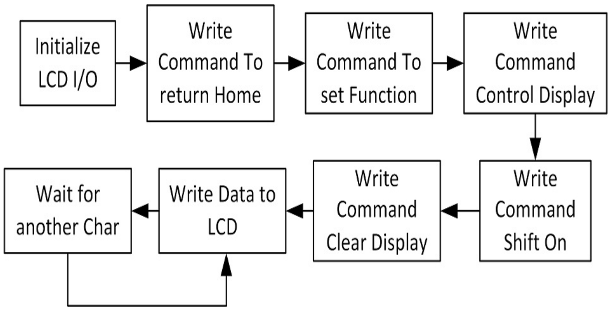

Initialize LCD I/O
Write Command To return Home
Write Command To set Function
Write Command Control Display
Write Command Shift On
Write Command Clear Display
Write Data to LCD
Wait for another Char

Figure 8: LCD Routine Flow Chart

E. Filter Block Implementation

The filter was a bit more complicated, it was decided to try to implement a filter that could work independently of the speed of the Admodel with a FIFO buffer. That idea later got scrapped for the final version that is a block with a 5 word wide circular buffer that updates with both channels from the Admodel and completing two filtrations each clock cycle. To be able to accomplish that with all the multiplications and additions a three stage pipeline was implemented to lower the load of the FPGA. Changing the coefficient was also implemented as parallel ports sent directly from the controller block.

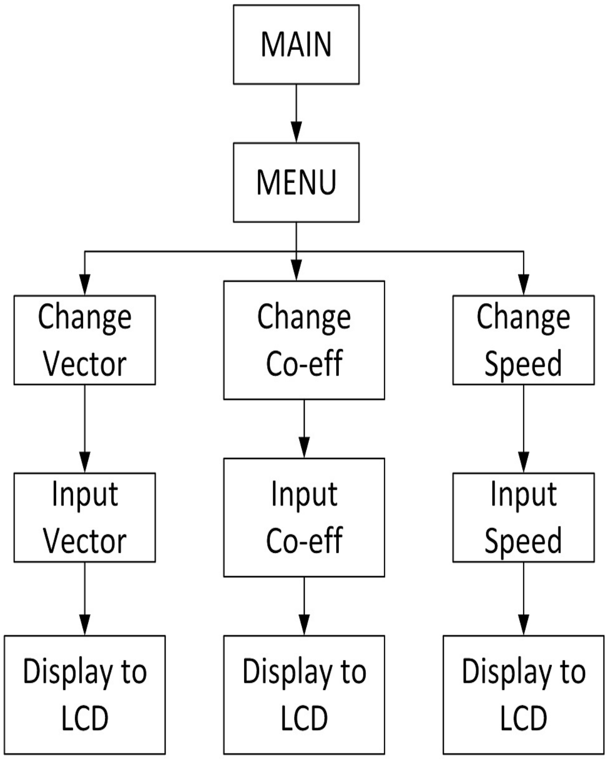


Figure 9: Main Interface Flow Chart

F. FindMax Block Implementation

The next stage in the chain was to find the position of the reference and echo pulse. As these operations are identical the block was implemented to run for 20 clock cycles send out the maximum $value(i)$, $value(i+1)$, $value(i-1)$ and the index to the distance block. When doing this it would start the distance and interpolation unit so it could work in parallel to findmax and do some precalculations as it was waiting for the value and position of the echo pulse. Findmax would then restart and try to find the echo pulse and then signal the distance unit again. Findmax is implemented as a simple comparator that first compare the two filtered values against each other and then to the current maximum value. If the value is larger than the last maximum the previous and next values are saved in a buffer until they are sent out. To be able to save the next value

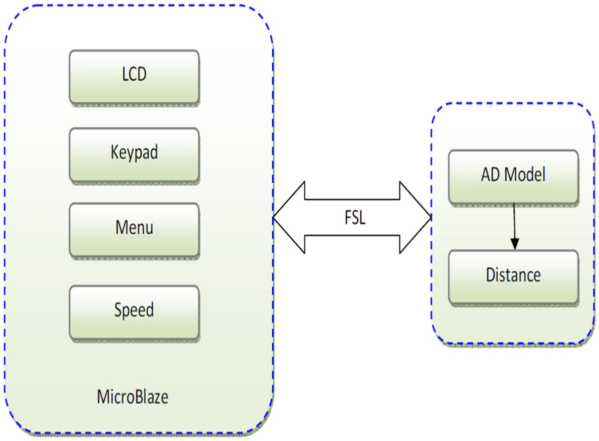


Figure 11: Block diagram of Mixed implementation

a flag was created to be able to save it the next clock cycle when it was fed from the Filter.

G. Distance and Interpolation

For both the reference and the echo, division has to be done to calculate the interpolation. The division is handled by a separate block that implements fast non-restoring division. The division can't handle negative numbers so the sign of the value b has to be checked and made to its two-complement and then the result from the division has to be made to its two complement to make the correct calculation when calculating j. This is done two times until both j_{ref} and j_{echo} are calculated.

The first is precalculated as mentioned above. Both of these distances have fractional numbers that is represented in fixed point with 5 bit precision. 5 bits were large enough that we would get the necessary precision for the 10 cm resolution that we need to display to the user, but at the same time not unnecessary large and take up space in the FPGA. As the speed of light and the sampling frequency in the last calculation is constant we precalculated this and implemented it as a constant inside the block. The result is then sent out through the FSL-link.

IV. Hardware Implementation

In complete hardware implementation phase a lot of things were already done, but it needed to be tied together with a controller and a menu system that could interface with all the other parts in VHDL. Modeling of speed were implemented, the keypad and LCD routines were ported to VHDL. The Fig. 12 shows the datapath and the different components of complete hardware implementation.

A. Distance Chain Implementation

The only changes here from what were used in mixed implementation, was the implementation of speed inside the Admodel. It was decided that the easiest way to do this would be to try to shift the echo pulse in each test vector, and when it is close to the end or reference, shift it the other way implying that the object were moving towards and away from the user. The speed of the simulation should be able to be set so we included a signal that was sent from the main program.

B. Main Block Implementation

The main program holds all the outside ports, all other components and the menu system. It has to listen for the key presses, start the distance chain, calculate speed and update variables when inside the menu. The main basically has two major components to keep track of, the distance calculation and the menu-system. When in distance calculation mode it starts the distance chain waits for it to finish and then depending on how long since it updated the LCD it might calculate speed and then update the LCD using DisplayText routine. If its in distance calculation mode, it can go into the menu by pressing ENTER.

C. DisplayText Routine Implementation

The main has to update variables and show these updated variables to the user together with a string that shows what the variable is i.e. (Testvector=variable), it was decided that sending whole strings to the LCD unit would take too much space so a system where a variable together with the screen (function) that the main want to write is sent. DisplayText routine takes the variable sent and converts it to a string using the component IntToString. When this is done a case statement chooses what to print out depending on the function. In each of these statements a custom string is built using some characters and the converted variable. This string is then fed one character at a time to the LCD and then printed. In the case of printing "D=distance S=speed" one extra variable is converted, for the integer value of the speed that comes from the speed block.

D. IntToString Routine Implementation

This block takes an integer between 0 and 999 and converts it to a three character string. If the number isn't 3 characters long it will be padded with spaces. The algorithm does checks against the size of the integer and then sends it to one of three different states: Ones, Tens and Hundreds.

E. Main Menu Implementation

The menu has to know what variable to update when PLUS or MINUS is pressed, and where to go whenever ENTER or BACK is pressed it also has to display this variable on the LCD. All variables is kept in an array and the whenever you move in the menu the index is updated and therefore the active variable that you can change is updated. Whenever a keypress happens the appropriate action is taken and then the LCD is updated by sending the variable and a function to DisplayText.

F. LCD Implementation

Hardware implementation of the LCD is almost a complete port from the C-code used previous two phases. When the device is set ON, the code initializes the LCD and then waits for DisplayText to send its first character. The limitations of the LCD regulate how long we have to set and hold the signals for it to register the change. This value has been set to 5ms using iterative methods.

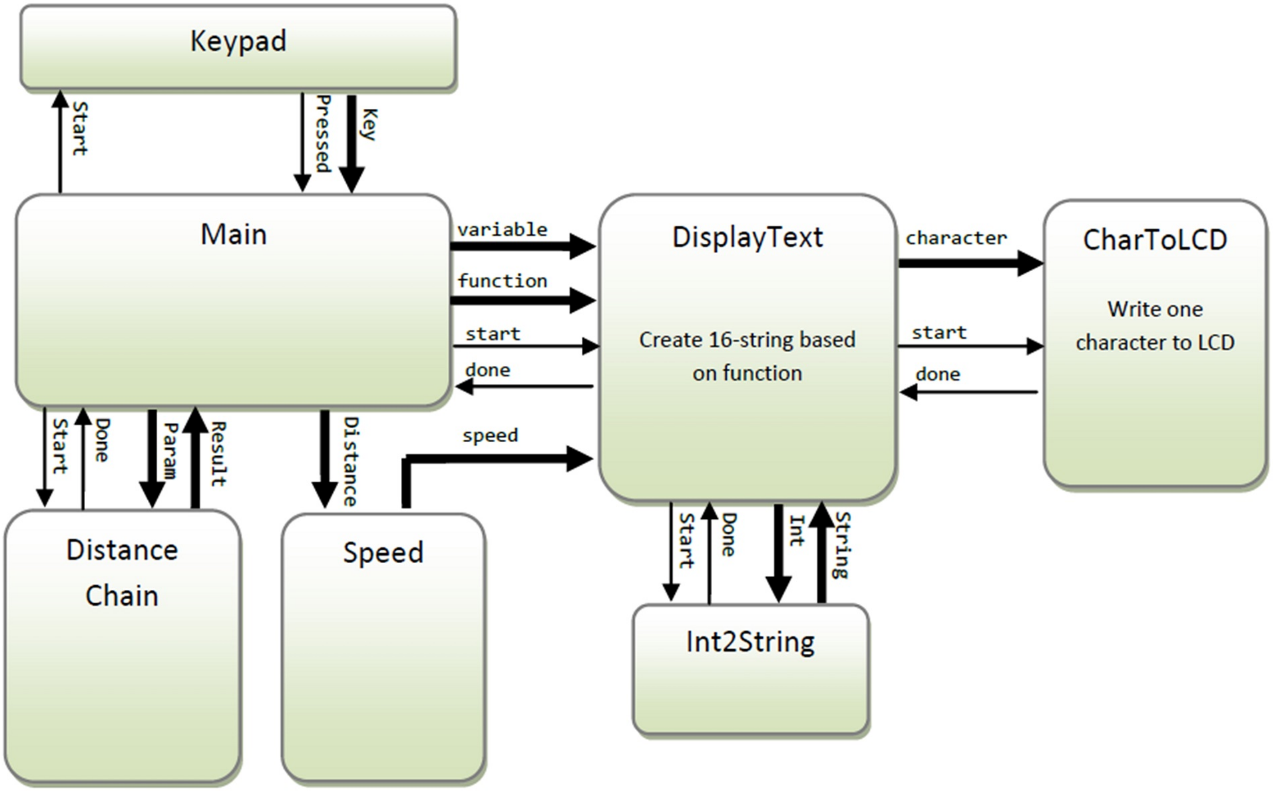


Figure 12: Overview of the datapath of distance chain for complete hardware implementation.

V. ASIC Implementation

The specification for the ASIC implementation was changed, following changes were made in the specification:

- Change in number of channels to one
- Change in data rate (200Mhz)
- Addition of SPI port
- Signed Coefficients
- Taps input Serially during initialization
- Result is in decimeter
- The data is input from pattern generator instead of AD model
- The output will be fed to logic analyzer

Fig. 13 is the Timing Diagram showing signals for all the major blocks of the system. An alternate design with a FIFO at the front end was also developed, distance chain itself was unchanged. The idea with this FIFO implementation was to allow the distance chain to run at a lower frequency and thus have lower power dissipation. Fig. 14 shows the block diagram of final ASIC design

A. FIR Filter Implementation

The filter is a 5 word buffer, on which the filter calculations are performed. To be able to accomplish all multiplications and additions required without increasing the performance load too much, a three stage pipeline has been introduced. The big changes in the filter block itself were the switch from two channels of data to one, and the use of signed coefficients. Where as the old filter block effectively had two filters, the new block only has one. The main addition to the filter block is the UART process which is used for reading coefficient data and feeding this to the filter.

B. Find Max Block Implementation

The next stage in the chain was to find the position of the reference and echo pulse. As these operations are identical the block was implemented to run for 20 clock cycles send out the maximum $value(i)$, $value(i+1)$, $value(i-1)$ and the index, to the distance block. When doing this it would start the distance and interpolation unit so it could work in parallel to findmax and do some pre-calculations as it was waiting for the value and position of the echo pulse. Findmax would then restart and try to find the echo pulse and then signal the distance unit again. Findmax is implemented as a simple comparator with a one word buffer. It checks the present value against the maximum value, if the present is higher it will store the present value and the value in the buffer, which is the previous value. It will also set a flag to store the next value. When the first 20 values have been checked the stored values will be sent to the distance unit, and the process will repeat for the echo values.

C. Distance and Interpolation

For both the reference and the echo, division has to be done to calculate the interpolation. The division is handled by

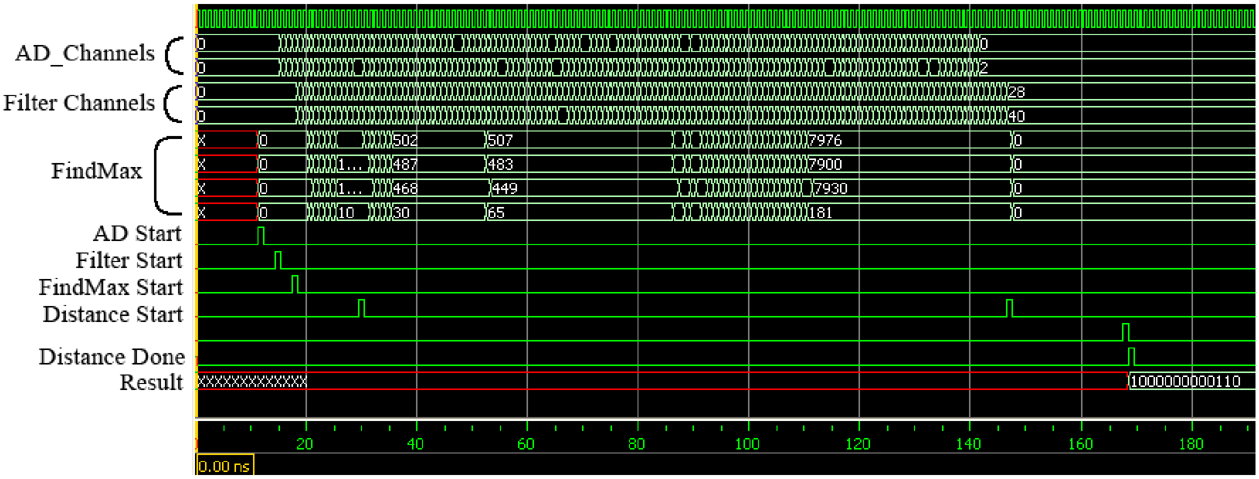


Figure 13: Timing Diagram showing signals for all the major blocks of the system

a separate block that implements fast non-restoring division. The division can't handle negative numbers so the sign of the value b has to be checked and made to its two-complement and then the result from the division has to be made to its two complement to make the correct calculation when calculating j. This is done two times until both j_{ref} and j_{echo} is calculated. The first is pre-calculated as mentioned above. As the speed of light and the sampling frequency in the last calculation is predefined we pre-calculated this and implemented it as a constant inside the block. The result is then sent out through the SPI.

D. Speed Calculation

The speed is calculated by storing a distance value, waiting a given time, then calculating the difference in distance compared to the stored value. This value is then multiplied with the waiting time divided by 1 second. For example the waiting time is by default set to 0.2s, to calculate the speed in dm/s the difference in distance is therefore multiplied by 5.

E. Serial Peripheral Interface (SPI) Implementation

An SPI port was implemented for outputting the result of measurements. The design uses this SPI master/slave interface, where the ASIC acts as master, for outputting of results. The SPI is designed according to the standard one-slave configuration, using mode 0. The SPI clock (SCLK) generated from the ASIC runs at 3.125 MHz. The SPI (in the ASIC) initiates a data transfer each 10 us, that is, at a frequency of 100 kHz. The data transfer is initiated by lowering the SS signal, according to the SPI standard. The output data is specified to be 12 bits unsigned values. The total number of bits for outputting over SPI is 16, for compliance with standard SPI components. The first bit is an indication of whether the data sent is a distance measurement or a speed measurement. A zero in this bit indicates distance, a one indicates speed.

F. Evaluation

There were two architectures which were considered. One was using FIFO so that the whole processing time can be spread over the whole spectrum of time. The FIFO was running on high speed clock i.e. 200 MHz where as rest of the design was working on slow clock i.e. 3.125 MHz. The other was running on the high speed clock. Synthesis of both the design was carried out at 4ns constraint to get a rough picture of area, timing, and power, table II shows the results.

Table II: FIFO and No FIFO

	FIFO	Without FIFO
Area [mm^2]	0.15	0.04
Timing [ns]	1.48	3.70
Power [mW]	18.54	6.68

Architecture without FIFO is taking 3 times less area then with FIFO. Both the designs fulfill the timing constraint of 4 ns but the design with FIFO is more efficient in terms of timing. The design with FIFO is 2.7 times expensive in term of power. The design without FIFO was chosen as it was less expensive in terms of power and area.

After synthesis power was simulated using VCD-files generated by the test bench, the table III show the results.

Table III: CLK Power

Net	Power(mW)	Cap(nF)
CLK	0.495	1.719

Using the expression $(P = f * Vdd^2 * C)$ clock power at 1.2V was 495072 nW, which verifies the result we got from the RTL compiler. The table IV shows power consumption in major blocks.

Table IV: Power consumption in major blocks

Block Name	Power (mW)
Filter	1.61
Distance	1.03
Findmax	0.88
SPI	0.09
Clkdiv	0.08

The results show that filter is the most power hungry

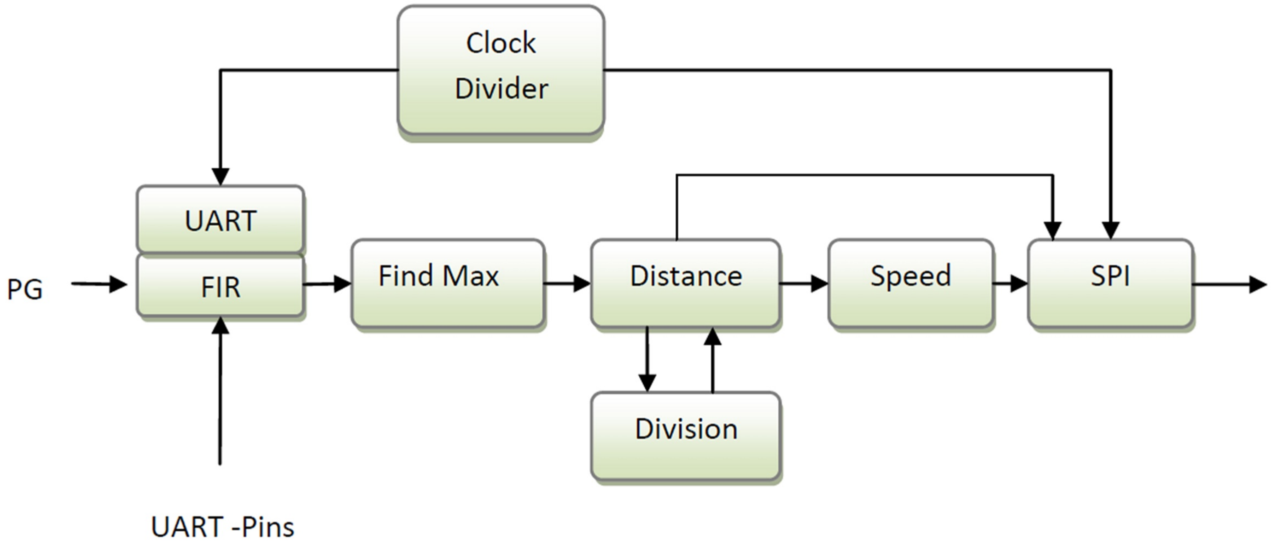


Figure 14: Overview of ASIC Design

block. The filter contains number of multipliers therefore it is consuming large part of power. The table V shows total power consumption.

Table V: Total Power

Leakage Power(mW)	Dynamic Power(mW)	Total Power(mW)
1.217	3.254	4.472

The table VI shows total area and timing results.

Table VI: Total Area and Timing

Area (um^2)	Timing (ns)
0.046	4.251

The timing constraint given was 5ns. The initial synthesis was carried out at medium effort and we got a slack of 749ps. The critical path was found to be between filter and findmax. The filter and findmax were consuming the major core area. Fig. 15 shows the layout of the chip.

Post layout the design was pad limited, using 24 pads, of which 3 were unused. The core area was $0.11mm^2$, with a core utilization of 43%. This gave a die area of $0.55mm^2$.

G. Verification

The complete design was verified with the help of test bench. This test bench would emulate AD converter functionality and send input to the design, and then it would receive distance and speed from the design and verify those results. The test bench is a comprehensive test bench designed to verify that the chip conformed to the specification with regard to functionality. The test bench is comprised of 14 test cases. Each test case is basically a component in the test bench, and they all share the component which holds the design. While a case is running it prints results to a corresponding text file, when it has finished it relinquishes control of the design to the next case. Test cases 1 through 8 tested distance measurements by moving the echo through the entire range of the specification, with different parameters. Test case 5 for example moves backwards through the range, and case 7 has negative coefficients. The remaining test cases were mainly for verification of speed measurements.

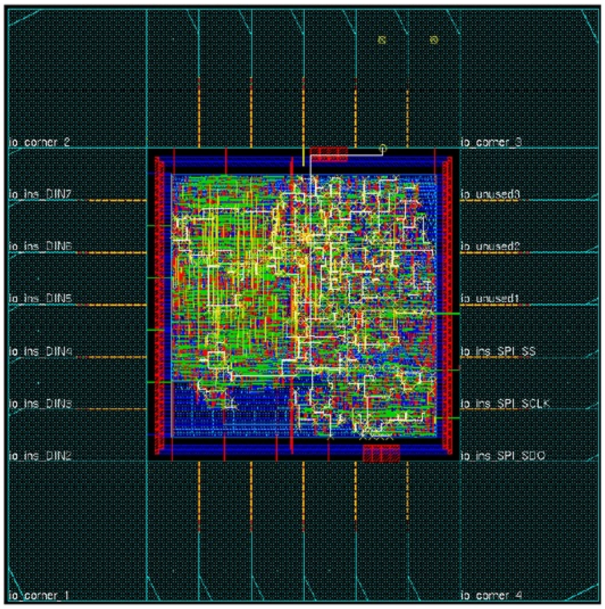


Figure 15: Layout of the chip

H. Physical Testing of Chip

The Fig. 16 shows pinout of the chip. For physical testing, power up the chip using 1.2V as VDD and Gnd as VSS. The interface has UART where you can feed co-efficient after reset, the coefficients are 8 bit wide and three coefficients needs to be entered. Put the DAV line high and feed 255 bytes to the chip. Put the DAV line low and you will get the result at SPI. The data out from the chip is 16 bit wide word and last three bits are redundant. The result of the speed calculation is after every 20,000 distance calculations. If the LSB is set the resulting word is speed and if it is clear the resulting word is distance.

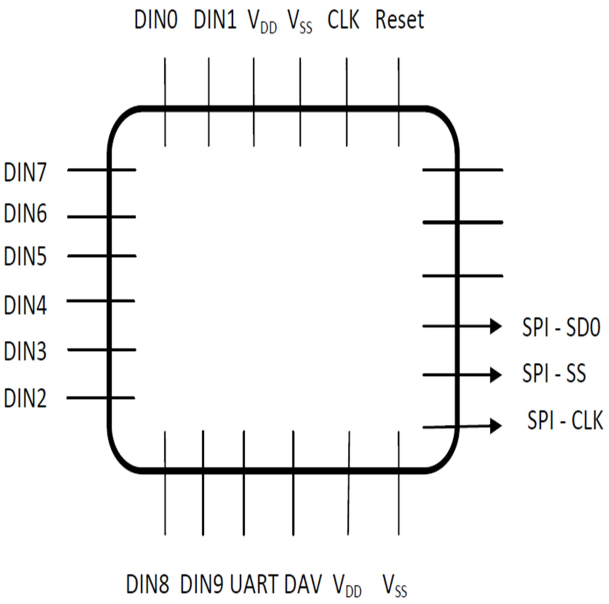


Figure 16: Pinout of the chip

VI. Conclusion

This paper presented a very practical problem and allowed the project group to approach it from several different viewpoints. This proved very useful as it showed the weaknesses and strengths of the platforms used in the project, or rather it showed how much impact these attributes have on real performance. For example it is quite expected that a hardware implementation should be faster than a software implementation. However, it was surprising to a roughly 400x increase in performance, which is the improvement from software phase to complete hardware implementation.

Acknowledgment

This work is partially supported by the Chinese Academic of Sciences (CAS) and The World Academy of Sciences (TWAS).

References

[1] T. Schlegl , T. Bretterklieber , M. Neumayer and H. Zangl “A novel sensor fusion concept for distance measurement in automotive applications”, IEEE Sensors, pp.775 -778 2010

[2] W. J. Fleming “New automotive sensors A review”, IEEE Sensors J., vol. 8, no. 11, pp.1900 -1921 2008

[3] T. Gandhi and M. M. Trivedi “Pedestrian protection systems: Issues, survey, and challenges”, IEEE Trans. Intell. Transp. Syst., vol. 8, no. 3, pp.413 -430 2007

[4] D. Marioli , C. Narduzzi , C. Offelli , D. Petri , E. Sardini and A. Taroni “Digital time-of-flight measurement for ultrasonic sensors”, IEEE Trans. Instrum. Meas., vol. 41, no. 1, pp.93 -97 1992

[5] C. Cai and P. P. L. Regtien “Accurate digital time-of-flight measurement using self-interference”, IEEE Trans. Instrum. Meas., vol. 42, no. 6, pp.990 -994 1993

[6] F. E. Gueuning , M. Varlan , C. E. Eugene and P. Dupuis “Accurate distance measurement by an autonomous ultrasonic system combining time-of-flight and phase-shift methods”, IEEE Trans. Instrum. Meas., vol. 46, no. 6, pp.1236 -1240 1997

[7] C. C. Tong , J. F. Figueroa and E. Barbieri “A method for short or long range time-of-flight measurements using phase-detection with an analog circuit”, IEEE Trans. Instrum. Meas., vol. 50, no. 5, pp.1324 -1328 2001

[8] S. S. Huang , C. F. Huang , K. N. Huang and M. S. Young “A high accuracy ultrasonic distance measurement system using binary frequency shift-keyed signal and phase detection”, Rev. Sci. Instrum., vol. 73, no. 10, pp.3671 -3677 2002

[9] L. Angrisani , A. Baccigalupi and R. S. L. Moriello “A measurement method based on Kalman filtering for ultrasonic time-of-flight estimation”, IEEE Trans. Instrum. Meas., vol. 55, no. 2, pp.442 -448 2006

[10] C. F. Huang , M. S. Young and Y. C. Li “Multiple-frequency continuous wave ultrasonic system for accurate distance measurement”, Rev. Sci. Instrum., vol. 70, no. 2, pp.1452 -1458 1999

[11] Y. S. Didosyan, H. Hauser, H. Wolfmayr, J. Nicolics and P. Fulmek “Magneto-optical rotational speed sensor”, Sens. Actuators A, Phys., vol. 106, no. 3, pp.168 -171 2003

[12] L. Wang, Y. Yan, Y. Hu and X. Qian “Rotational speed measurement through electrostatic sensing and correlation signal processing”, IEEE Trans. Instrum. Meas., vol. 63, no. 5, pp.1190 -1199 2014

[13] Y. Yan, B. Byrne, S. Woodhead and J. Coulthard “Velocity measurement of pneumatically conveyed solids using electrodynamic sensors”, Meas. Sci. Technol., vol. 6, no. 5, pp.515 -537 1995

[14] J. Ma and Y. Yan “Design and evaluation of electrostatic sensors for the measurement of velocity of pneumatically conveyed solids”, Flow Meas. Instrum., vol. 11, no. 3, pp.195-204 2000

[15] Lijuan Wang, Yong Yan , Yonghui Hu , Xiangchen Qian “Rotational Speed Measurement Using Single and Dual Electrostatic Sensors”, IEEE Sensors, pp. 1784-1793 2014

[16] Xilinx Inc. FPGA Design Tools. Silicon Devices. www.xilinx.com

[17] Xilinx. MicroBlaze. www.xilinx.com/tools/microblaze.htm

[18] Xilinx Virtex-II Board. www.xilinx.com/univ/xupv2p.html

[19] Xilinx FSL. www.xilinx.com/products/intellectual-property/fsl.html

[20] STMicroelectronics. www.st.com/web/en/home.html

[21] Cadence EDA Tools. www.cadence.com/en/default.aspx

[22] M. Imai, “Embedded tutorial: hardware/software codesign”, IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 1999.

[23] J. Noguera, R.M. Badia, “HW/SW codesign techniques for dynamically reconfigurable architectures” IEEE Trans. Very Large Scale Integration (VLSI) Systems, vol. 10, no. 4, pp. 399-415, Aug. 2002.

[24] M. D. Edwards, et al., “Acceleration of software algorithms using hardware/software co-design techniques”, J. Syst. Architecture, vol. 42, no. 9/10, pp.1997.

[25] W. Wolf, “A Decade of Hardware/Software Codesign,” IEEE Computer, vol. 36, pp. 38-43, April 2003

[26] R. Ernst, J. Henkel, and T. Benner, “Hardware-Software Cosynthesis for Microcontrollers,” IEEE Transaction on Design and Test of Computers, vol. 10, pp. 64-75, December 1993

[27] W. Wolf, “Hardware/software Co-design of Embedded Systems,” IEEE Proceeding, vol. 82, pp. 967-989, July 1994

[28] Y. Li, T. Callahan, E. Darnell, R. Harr, U. Kurkure, and J. Stockwood, “Hardware-Software Co-Design of Embedded Reconfigurable Architectures,” in Proceeding of 37th Design Automation Conference, pp. 507-512, June 2000

[29] Vermeulen, L. Nachtergaele, F. Catthoor, D. Verkest, and H. De Man, “Flexible Hardware Acceleration for Multimedia Oriented Microprocessors,” IEEE Transactions on Very Large Scale Integration Systems, pp. 171-177, December 2000

[30] M. Boden, J. Schneider, K. Feske, and S. Rulke, “Enhanced Reusability for SoC-based HW/SW Co-design,” in Proceeding of Euromicro Symposium on Digital System Design 2002, pp. 94-99, September 2002.

[31] Kai-Yuan Jan, Chih-Bin Fan, An-Chao Kuo, Wen-Chi Yen, and Youn-Long Lin, "A Platform-based SOC Design Methodology and Its Application in Image Compression," Special Issue on HW-SW Codesign for SoC, International Journal of Embedded Systems, Inderscience Publishers, USA. Vol. 1, Issue 1/2, pp. 23-32, 2005.

[32] Chiodo, M. and et al. Hardware-software codesign of embedded systems. In IEEE Micro, 1994.

[33] Ernst, R. and et al. Codesgin of embedded systems: status and trends. In Proceedings of IEEE Design and Test of Computers, 1998.

[34] Gallery, R. and et al. Hardware/software partitioning and simulation with SystemC. In Proceedings of the 2nd WSEAS ICECSP, 2003.

[35] Hurk, J. and et al. System Level Hardware/Software Co-Design: An Industrial Approach, 1997.

[36] De Micheli, G. and et al. Hardware/Software Co-design. In Proceedings of the IEEE, 1997.

On the Codes over a Semilocal Finite Ring

Abdullah Dertli
Department of Mathematics
Ondokuz Mayıs University
Samsun, Turkey

Yasemin Cengellenmis
Department of Mathematics
Trakya University
Edirne, Turkey

Senol Eren
Department of Mathematics
Ondokuz Mayıs University
Samsun, Turkey

***Abstract*—In this paper, we study the structure of cyclic, quasi cyclic, constacyclic codes and their skew codes over the finite ring R. The Gray images of cyclic, quasi cyclic, skew cyclic, skew quasi cyclic and skew constacyclic codes over R are obtained. A necessary and sufficient condition for cyclic (negacyclic) codes over R that contains its dual has been given. The parameters of quantum error correcting codes are obtained from both cyclic and negacyclic codes over R. Some examples are given. Firstly, quasi constacyclic and skew quasi constacyclic codes are introduced. By giving two inner product, it is investigated their duality. A sufficient condition for 1 generator skew quasi constacyclic codes to be free is determined.**

Keywords*—*Cyclic codes; Skew cyclic codes; Quantum codes

I. Introduction

In the beginning, a lot of research on error-correcting codes are concentrated on codes over finite fields. Since the revelation in 1994 [17], there has been a lot of interest in codes over finite rings. The structure of a certain type of codes over many rings are determined such as negacyclic, cyclic, quasi-cyclic, consta cyclic codes in [6,11,20,21,22,23,26,32]. Many methods and many approaches are applied to produce certain types of codes with good parameters and properties.

Some authors generalized the notion of cyclic, quasi-cyclic and constacyclic codes by using generator polynomials in skew polynomial rings [1,2,5,7,8,9,14,15,18,27,30].

Moreover, in [10] Calderbank et al. gave a way to construct quantum error correcting codes from the classical error-correcting codes, although the theory of quantum error-correcting codes has striking differences from the theory of classical error correcting codes. Many good quantum codes have been constructed by using classical cyclic codes over finite fields or finite rings with self orthogonal (or dual containing) properties in [3,12,13,16,19,24,25,28,29,31].

In [4] they introduced the finite ring $R = Z_3[v]/\left\langle v^3 - v\right\rangle$. They studied the structure of this ring. The algebraic structure of cyclic and dual codes was also studied. A MacWilliams type identity was established.

In this paper, first of all we gave some definitions. By giving the duality of codes via inner product, it is shown that C is self orthogonal code over R, so is $\phi(C)$, where ϕ is a Gray map.

The Gray images of cyclic and quasi-cyclic codes over R are obtained. A linear code over R is represented using three ternary codes and the generator matrix is given.

After a cyclic (negacyclic) code over R is represented via cyclic (negacyclic) codes over Z_3, it is determined the dual of cyclic (negacyclic) code. A necessary and sufficient condition for cyclic (negacyclic) code over R that contains its dual is given. The parameters of quantum error-correcting codes are obtained from both cyclic and negacyclic codes over R. As a last, some examples are given about quantum error-correcting codes.

When n is odd, it is defined the λ-constacyclic codes over R where λ is unit. A constacyclic code is represented using either cyclic codes or negacyclic codes of length n.

It is found the nontrivial automorphism θ on the ring R. By using this automorphism, the skew cyclic, skew quasi-cyclic and skew constacyclic codes over R are introduced. The number of distinct skew cyclic codes over R is given. The Gray images of skew codes are obtained.

Firstly, quasi-constacyclic and skew quasi-constacyclic codes over R are introduced. By using two inner product, it is investigated the duality about quasi-constacyclic and skew quasi-constacyclic codes over R. The Gray image of skew quasi-constacyclic codes over R is determined. A sufficient condition for 1-generator skew quasi-constacyclic code to be free is determined.

II. Preliminaries

Suppose $R = Z_3 + vZ_3 + v^2Z_3$ where $v^3 = v$ and $Z_3 = \{0, 1, 2\}$. R is a finite commutative ring with 27 elements. This ring is a semi local ring with three maximal ideals. R is a principal ideal ring and not finite chain ring. The units of the ring are $1, 2, 1+v^2, 1+v+2v^2, 1+2v+2v^2, 2+v+v^2, 2+2v+v^2, 2+2v^2$. The maximal ideals,

$$\begin{aligned}
\langle v\rangle &= \langle 2v\rangle = \left\langle v^2\right\rangle = \left\langle 2v^2\right\rangle \\
&= \{0, v, 2v, v^2, 2v^2, v+v^2, v+2v^2, 2v+v^2, \\
&\quad 2v+2v^2\} \\
\langle 1+v\rangle &= \langle 2+2v\rangle = \left\langle 1+2v+v^2\right\rangle = \left\langle 2+v+2v^2\right\rangle \\
&= \{0, 1+v, 2+2v, v+v^2, 2v+2v^2, 1+2v \\
&\quad +v^2, 1+2v^2, 2+v^2, 2+v+2v^2\} \\
\left\langle 1+v+v^2\right\rangle &= \langle 1+2v\rangle = \langle 2+v\rangle = \left\langle 2+2v+2v^2\right\rangle \\
&= \{0, 2+v, 1+2v, 2v+v^2, v+2v^2, 2+v^2, \\
&\quad 1+2v^2, 2+2v+2v^2, 1+v+v^2\}
\end{aligned}$$

The other ideals,

$$\begin{aligned} \langle 0\rangle &= \{0\}\\ \langle 1\rangle &= \langle 2\rangle = \left\langle 1+v^2\right\rangle = \left\langle 1+v+2v^2\right\rangle \\ &= \left\langle 1+2v+2v^2\right\rangle = \left\langle 2+v+v^2\right\rangle \\ &= \left\langle 2+2v+v^2\right\rangle = \left\langle 2+2v^2\right\rangle = R\\ \left\langle 1+2v^2\right\rangle &= \left\langle 2+v^2\right\rangle = \{0,2+v^2,1+2v^2\}\\ \left\langle v+v^2\right\rangle &= \left\langle 2v+2v^2\right\rangle = \{0,v+v^2,2v+2v^2\}\\ \left\langle v+2v^2\right\rangle &= \left\langle 2v+v^2\right\rangle = \{0,v+2v^2,2v+v^2\}\end{aligned}$$

A linear code C over R length n is a $R-$submodule of R^n. An element of C is called a codeword.

For any $x=(x_0,x_1,...,x_{n-1})$, $y=(y_0,y_1,...,y_{n-1})$ the inner product is defined as

$$x.y=\sum_{i=0}^{n-1} x_i y_i$$

If $x.y=0$ then x and y are said to be orthogonal. Let C be linear code of length n over R, the dual code of C

$$C^{\perp}=\{x:\forall y\in C, x.y=0\}$$

which is also a linear code over R of length n. A code C is self orthogonal if $C\subseteq C^{\perp}$ and self dual if $C=C^{\perp}$.

A cyclic code C over R is a linear code with the property that if $c=(c_0,c_1,...,c_{n-1})\in C$ then $\sigma(C)=(c_{n-1},c_0,...,c_{n-2})\in C$. A subset C of R^n is a linear cyclic code of length n iff it is polynomial representation is an ideal of $R[x]/\langle x^n-1\rangle$.

A constacyclic code C over R is a linear code with the property that if $c=(c_0,c_1,...,c_{n-1})\in C$ then $\nu(C)=(\lambda c_{n-1},c_0,...,c_{n-2})\in C$ where λ is a unit element of R. A subset C of R^n is a linear λ-constacyclic code of length n iff it is polynomial representation is an ideal of $R[x]/\langle x^n-\lambda\rangle$.

A negacyclic code C over R is a linear code with the property that if $c=(c_0,c_1,...,c_{n-1})\in C$ then $\eta(C)=(-c_{n-1},c_0,...,c_{n-2})\in C$. A subset C of R^n is a linear negacyclic code of length n iff it is polynomial representation is an ideal of $R[x]/\langle x^n+1\rangle$.

Let C be code over Z_3 of length n and $\acute{c}=(\acute{c}_0,\acute{c}_1,...,\acute{c}_{n-1})$ be a codeword of C. The Hamming weight of $\acute{c}$ is defined as $w_H(\acute{c})=\sum_{i=0}^{n-1} w_H(\acute{c}_i)$ where $w_H(\acute{c}_i)=1$ if $\acute{c}_i\neq 0$ and $w_H(\acute{c}_i)=0$ if $\acute{c}_i=0$. Hamming distance of C is defined as $d_H(C)=\min d_H(c,\acute{c})$, where for any $\acute{c}\in C$, $c\neq\acute{c}$ and $d_H(c,\acute{c})$ is Hamming distance between two codewords with $d_H(c,\acute{c})=w_H(c-\acute{c})$.

Let $a\in Z_3^{3n}$ with $a=(a_0,a_1,...,a_{3n-1})=\left(a^{(0)}\left|a^{(1)}\right|a^{(2)}\right)$, $a^{(i)}\in Z_3^n$ for $i=0,1,2$. Let φ be a map from Z_3^{3n} to Z_3^{3n} given by $\varphi(a)=\left(\sigma\left(a^{(0)}\right)\left|\sigma\left(a^{(1)}\right)\right|\sigma\left(a^{(2)}\right)\right)$ where σ is a cyclic shift from Z_3^n to Z_3^n given by $\sigma\left(a^{(i)}\right)=((a^{(i,n-1)}),(a^{(i,0)}),(a^{(i,1)}),...,(a^{(i,n-2)}))$ for every $a^{(i)}=(a^{(i,0)},...,a^{(i,n-1)})$ where $a^{(i,j)}\in Z_3$, $j=0,1,...,n-1$. A code of length $3n$ over Z_3 is said to be quasi cyclic code of index 3 if $\varphi(C)=C$.

Let $n=sl$. A quasi-cyclic code C over R of length n and index l is a linear code with the property that if

$e=(e_{0,0},...,e_{0,l-1},e_{1,0},...,e_{1,l-1},...,e_{s-1,0},...,e_{s-1,l-1})\in C$, then $\tau_{s,l}(e)=(e_{s-1,0},...,e_{s-1,l-1},e_{0,0},...,e_{0,l-1},...,e_{s-2,0},...,e_{s-2,l-1})\in C$.

Let $a\in Z_3^{3n}$ with $a=(a_0,a_1,...,a_{3n-1})=\left(a^{(0)}\left|a^{(1)}\right|a^{(2)}\right)$, $a^{(i)}\in Z_3^n$, for $i=0,1,2$. Let Γ be a map from Z_3^{3n} to Z_3^{3n} given by

$$\Gamma(a)=\left(\mu\left(a^{(0)}\right)\left|\mu\left(a^{(1)}\right)\right|\mu\left(a^{(2)}\right)\right)$$

where μ is the map from Z_3^n to Z_3^n given by

$$\mu\left(a^{(i)}\right)=((a^{(i,s-1)}),(a^{(i,0)}),...,(a^{(i,s-2)}))$$

for every $a^{(i)}=\left(a^{(i,0)},...,a^{(i,s-1)}\right)$ where $a^{(i,j)}\in Z_3^l$, $j=0,1,...,s-1$ and $n=sl$. A code of length $3n$ over Z_3 is said to be $l-$quasi cyclic code of index 3 if $\Gamma(C)=C$.

III. Gray Map and Gray Images of Cyclic and Quasi-cyclic Codes Over R

In [4], the Gray map is defined as follows

$$\begin{aligned}\phi &: R\rightarrow Z_3^3\\ \phi(a+vb+v^2c) &= (a,a+b+c,a+2b+c)\end{aligned}$$

Let C be a linear code over R of length n. For any codeword $c=(c_0,...,c_{n-1})$ the Lee weight of c is defined as $w_L(c)=\sum_{i=0}^{n-1} w_L(c_i)$ and the Lee distance of C is defined as $d_L(C)=\min d_L(c,\acute{c})$, where for any $\acute{c}\in C$, $c\neq\acute{c}$ and $d_L(c,\acute{c})$ is Lee distance between two codewords with $d_L(c,\acute{c})=w_L(c-\acute{c})$. Gray map ϕ can be extended to map from R^n to Z_3^{3n}.

Theorem 1: The Gray map ϕ is a weight preserving map from $(R^n,\text{Lee weight})$ to $\left(Z_3^{3n},\text{Hamming weight}\right)$. Moreover it is an isometry from R^n to Z_3^{3n}.

Theorem 2: If C is an $[n,k,d_L]$ linear codes over R then $\phi(C)$ is a $[3n,k,d_H]$ linear codes over Z_3, where $d_H=d_L$.

Proof: Let $x=a_1+vb_1+v^2c_1$, $y=a_2+vb_2+v^2c_2\in R,\alpha\in Z_3$ then

$\phi(x+y)=\phi\left(a_1+a_2+v(b_1+b_2)+v^2(c_1+c_2)\right)$

$=(a_1+a_2,a_1+a_2+b_1+b_2+c_1+c_2,a_1+a_2+2(b_1+b_2)+c_1+c_2)$

$=(a_1,a_1+b_1+c_1,a_1+2b_1+c_1)+(a_2,a_2+b_2+c_2,a_2+2b_2+c_2)$

$=\phi(x)+\phi(y)$

$\phi(\alpha x)=\phi\left(\alpha a_1+v\alpha b_1+v^2\alpha c_1\right)$

$=(\alpha a_1,\alpha a_1+\alpha b_1+\alpha c_1,\alpha a_1+2\alpha b_1+\alpha c_1)$

$=\alpha(a_1,a_1+b_1+c_1,a_1+2b_1+c_1)$

$=\alpha\phi(x)$

so ϕ is linear. As ϕ is bijective then $|C|=|\phi(C)|$. From Theorem 1 we have $d_H=d_L$. ■

Theorem 3: If C is self orthogonal, so is $\phi(C)$.

Proof: Let $x = a_1 + vb_1 + v^2c_1,\ y = a_2 + vb_2 + v^2c_2$ where $a_1, b_1, c_1,\ a_2, b_2, c_2 \in Z_3$. From $x.y = a_1a_2 + v(a_1b_2 + b_1a_2 + b_1c_2 + c_1b_2) + v^2(a_1c_2 + b_1b_2 + c_1a_2 + c_1c_2)$ if C is self orthogonal, so we have

$$a_1a_2 = 0,$$
$$a_1b_2 + b_1a_2 + b_1c_2 + c_1b_2 = 0,$$
$$a_1c_2 + b_1b_2 + c_1a_2 + c_1c_2 = 0.$$

From $\phi(x).\phi(y) = (a_1, a_1 + b_1 + c_1, a_1 + 2b_1 + c_1)(a_2, a_2 + b_2 + c_2, a_2 + 2b_2 + c_2) = a_1a_2 + a_1a_2 + a_1b_2 + a_1c_2 + b_1a_2 + b_1b_2 + b_1c_2 + c_1a_2 + c_1b_2 + c_1c_2 + a_1a_2 + 2(a_1b_2 + b_1a_2 + b_1c_2 + c_1b_2) + a_1c_2 + b_1b_2 + c_1a_2 + c_1c_2 = 0$
Therefore, we have $\phi(C)$ is self orthogonal. ■

Note that $\phi(C)^{\perp} = \phi\left(C^{\perp}\right)$. Moreover, if C is self-dual, so is $\phi(C)$.

Proposition 4: Let ϕ the Gray map from R^n to Z_3^{3n}, let σ be cyclic shift and let φ be a map as in the preliminaries. Then $\phi\sigma = \varphi\phi$.

Proof: Let $r_i = a_i + vb_i + v^2c_i$ be the elements of R for $i = 0, 1, ..., n-1$. We have $\sigma(r_0, r_1, ..., r_{n-1}) = (r_{n-1}, r_0, ..., r_{n-2})$. If we apply ϕ, we have

$$\begin{aligned}\phi(\sigma(r_0, ..., r_{n-1})) &= \phi(r_{n-1}, r_0, ..., r_{n-2}) \\ &= (a_{n-1}, ..., a_{n-2}, a_{n-1} + b_{n-1} + c_{n-1} \\ &\quad , ..., a_{n-2} + b_{n-2} + c_{n-2}, a_{n-1} + \\ &\quad 2b_{n-1} + c_{n-1}, ..., a_{n-2} + 2b_{n-2} + \\ &\quad c_{n-2})\end{aligned}$$

On the other hand $\phi(r_0, ..., r_{n-1}) = (a_0, ..., a_{n-1}, a_0 + b_0 + c_0, ..., a_{n-1} + b_{n-1} + c_{n-1}, a_0 + 2b_0 + c_0, ..., a_{n-1} + 2b_{n-1} + c_{n-1})$. If we apply φ, we have $\varphi(\phi(r_0, r_1, ..., r_{n-1})) = (a_{n-1}, ..., a_{n-2}, a_{n-1} + b_{n-1} + c_{n-1}, ..., a_{n-2} + b_{n-2} + c_{n-2}, a_{n-1} + 2b_{n-1} + c_{n-1}, ..., a_{n-2} + 2b_{n-2} + c_{n-2})$. Thus, $\phi\sigma = \varphi\phi$. ■

Proposition 5: Let σ and φ be as in the preliminaries. A code C of length n over R is cyclic code if and only if $\phi(C)$ is quasi cyclic code of index 3 over Z_3 with length $3n$.

Proof: Suppose C is cyclic code. Then $\sigma(C) = C$. If we apply ϕ, we have $\phi(\sigma(C)) = \phi(C)$. From Proposition 4, $\phi(\sigma(C)) = \varphi(\phi(C)) = \phi(C)$. Hence, $\phi(C)$ is a quasi cyclic code of index 3. Conversely, if $\phi(C)$ is a quasi cyclic code of index 3, then $\varphi(\phi(C)) = \phi(C)$. From Proposition 4, we have $\varphi(\phi(C)) = \phi(\sigma(C)) = \phi(C)$. Since ϕ is injective, it follows that $\sigma(C) = C$. ■

Proposition 6: Let $\tau_{s,l}$ be quasi-cyclic shift on R. Let Γ be as in the preliminaries. Then $\phi\tau_{s,l} = \Gamma\phi$.

Proof: Let $e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1})$ with $e_{i,j} = a_{i,j} + vb_{i,j} + v^2c_{i,j}$ where $i = 0, 1, ..., s-1$ and $j = 0, 1, ..., l-1$. We have $\tau_{s,l}(e) = (e_{s-1,0}, ..., e_{s-1,l-1}, e_{0,0}, ..., e_{0,l-1}, ..., e_{s-2,0}, ..., e_{s-2,l-1})$. If we apply ϕ, we have

$$\begin{aligned}\phi(\tau_{s,l}(e)) &= (a_{s-1,0}, ..., a_{s-2,l-1}, a_{s-1,0} + b_{s-1,0} + c_{s-1,0} \\ &\quad , ..., a_{s-2,l-1} + b_{s-2,l-1} + c_{s-2,l-1}, a_{s-1,0} + \\ &\quad 2b_{s-1,0} + c_{s-1,0}, ..., a_{s-2,l-1} + 2b_{s-2,l-1} + \\ &\quad c_{s-2,l-1})\end{aligned}$$

On the other hand,

$$\begin{aligned}\phi(e) &= (a_{0,0}, ..., a_{s-1,l-1}, a_{0,0} + b_{0,0} + c_{0,0}, ..., a_{s-1,l-1} + \\ &\quad b_{s-1,l-1} + c_{s-1,l-1}, a_{0,0} + 2b_{0,0} + c_{0,0}, ..., a_{s-1,l-1} \\ &\quad +2b_{s-1,l-1} + c_{s-1,l-1})\end{aligned}$$

$\Gamma(\varphi(e)) = (a_{s-1,0}, ..., a_{s-2,l-1}, a_{s-1,0} + b_{s-1,0} + c_{s-1,0}, ..., a_{s-2,l-1} + b_{s-2,l-1} + c_{s-2,l-1}, a_{s-1,0} + 2b_{s-1,0} + c_{s-1,0}, ..., a_{s-2,l-1} + 2b_{s-2,l-1} + c_{s-2,l-1})$. So, we have $\varphi\tau_{s,l} = \Gamma\varphi$. ■

Theorem 7: The Gray image of a quasi-cyclic code over R of length n with index l is a l-quasi cyclic code of index 3 over Z_3 with length $3n$.

Proof: Let C be a quasi-cyclic code over R of length n with index l. That is $\tau_{s,l}(C) = C$. If we apply ϕ, we have $\phi(\tau_{s,l}(C)) = \phi(C)$. From the Proposition 6, $\phi(\tau_{s,l}(C)) = \phi(C) = \Gamma(\phi(C))$. So, $\phi(C)$ is a l quasi-cyclic code of index 3 over Z_3 with length $3n$. ■

We denote that $A_1 \otimes A_2 \otimes A_3 = \{(a_1, a_2, a_3) : a_1 \in A_1, a_2 \in A_2, a_3 \in A_3\}$ and $A_1 \oplus A_2 \oplus A_3 = \{a_1 + a_2 + a_3 : a_1 \in A_1, a_2 \in A_2, a_3 \in A_3\}$

Let C be a linear code of length n over R. Define

$$\begin{aligned}C_1 &= \left\{a \in Z_3^n : \exists b, c \in Z_3^n, a + vb + v^2c \in C\right\} \\ C_2 &= \left\{a + b + c \in Z_3^n : a + vb + v^2c \in C\right\} \\ C_3 &= \left\{a + 2b + c \in Z_3^n : a + vb + v^2c \in C\right\}\end{aligned}$$

Then C_1, C_2 and C_3 are ternary linear codes of length n. Moreover, the linear code C of length n over R can be uniquely expressed as $C = (1 + 2v^2)C_1 \oplus \left(2v + 2v^2\right)C_2 \oplus \left(v + 2v^2\right)C_3$.

Theorem 8: Let C be a linear code of length n over R. Then $\phi(C) = C_1 \otimes C_2 \otimes C_3$ and $|C| = |C_1|\,|C_2|\,|C_3|$.

Proof: For any $(a_0, a_1, ..., a_{n-1}, a_0 + b_0 + c_0, a_1 + b_1 + c_1, ..., a_{n-1} + b_{n-1} + c_{n-1}, a_0 + 2b_0 + c_0, a_1 + 2b_1 + c_1, ..., a_{n-1} + 2b_{n-1} + c_{n-1}) \in \phi(C)$. Let $m_i = a_i + vb_i + v^2c_i,\ i = 0, 1, ..., n-1$. Since ϕ is a bijection $m = (m_0, m_1, ..., m_{n-1}) \in C$. By definitions of C_1, C_2 and C_3 we have $(a_0, a_1, ..., a_{n-1}) \in C_1, (a_0 + b_0 + c_0, a_1 + b_1 + c_1, ..., a_{n-1} + b_{n-1} + c_{n-1}) \in C_2, (a_0 + 2b_0 + c_0, a_1 + 2b_1 + c_1, ..., a_{n-1} + 2b_{n-1} + c_{n-1}) \in C_3$. So, $(a_0, a_1, ..., a_{n-1}, a_0 + b_0 + c_0, a_1 + b_1 + c_1, ..., a_{n-1} + b_{n-1} + c_{n-1}, a_0 + 2b_0 + c_0, a_1 + 2b_1 + c_1, ..., a_{n-1} + 2b_{n-1} + c_{n-1}) \in C_1 \otimes C_2 \otimes C_3$. That is $\phi(C) \subseteq C_1 \otimes C_2 \otimes C_3$.

On the other hand, for any $(a, b, c) \in C_1 \otimes C_2 \otimes C_3$ where $a = (a_0, a_1, ..., a_{n-1}) \in C_1$, $b = (a_0 + b_0 + c_0, a_1 + b_1 + c_1, ..., a_{n-1} + b_{n-1} + c_{n-1}) \in C_2$, $c = (a_0 + 2b_0 + c_0, a_1 + 2b_1 + c_1, ..., a_{n-1} + 2b_{n-1} + c_{n-1}) \in C_3$. There are $x = (x_0, x_1, ..., x_{n-1})$, $y = (y_0, y_1, ..., y_{n-1})$, $z = (z_0, z_1, ..., z_{n-1}) \in C$ such that $x_i = a_i + (v + 2v^2)p_i$, $y_i = b_i + \left(1 + 2v^2\right)q_i$, $z_i = c_i + \left(2v + 2v^2\right)r_i$ where $p_i,\ q_i, r_i \in Z_3$ and $0 \leq i \leq n-1$. Since C is linear we have $m = (1 + 2v^2)x + \left(2v + 2v^2\right)y + \left(v + 2v^2\right)z = a + v(2b + c) + v^2(2a + 2b + 2c) \in C$. It follows then $\phi(m) = (a, b, c)$, which gives $C_1 \otimes C_2 \otimes C_3 \subseteq \phi(C)$.

Therefore, $\phi(C) = C_1 \otimes C_2 \otimes C_3$. The second result is easy to verify. ■

Corollary 9: If $\phi\left(C\right) = C_1 \otimes C_2 \otimes C_3$, then $C = (1 + 2v^2)C_1 \oplus \left(2v + 2v^2\right) C_2 \oplus \left(v + 2v^2\right) C_3$. It is easy to see that

$$|C| = |C_1|\,|C_2|\,|C_3| = 3^{n-\deg(f_1)} 3^{n-\deg(f_2)} 3^{n-\deg(f_3)}$$
$$= 3^{3n-(\deg(f_1)+\deg(f_2)+\deg(f_3))}$$

where f_1, f_2 and f_3 are the generator polynomials of C_1, C_2 and C_3, respectively.

Corollary 10: If G_1, G_2 and G_3 are generator matrices of ternary linear codes C_1, C_2 and C_3 respectively, then the generator matrix of C is

$$G = \begin{bmatrix} (1+2v^2)G_1 \\ \left(2v+2v^2\right) G_2 \\ \left(v+2v^2\right) G_3 \end{bmatrix}.$$

We have

$$\phi(G) = \begin{bmatrix} \phi((1+2v^2)G_1) \\ \phi(\left(2v+2v^2\right) G_2) \\ \phi(\left(v+2v^2\right) G_3) \end{bmatrix} = \begin{bmatrix} G_1 & 0 & 0 \\ 0 & G_2 & 0 \\ 0 & 0 & G_3 \end{bmatrix}.$$

Let d_L minimum Lee weight of linear code C over R. Then, $d_L = d_H(\phi\left(C\right)) = \min\{d_H(C_1), d_H(C_2), d_H(C_3)\}$ where $d_H(C_i)$ denotes the minimum Hamming weights of ternary codes C_1, C_2 and C_3, respectively.

As similiar to section 4 in [4] we have the following Lemma and Examples.

Lemma 11: Let $C = \langle f(x)\rangle$ be a negacyclic code of length n over R and $\phi\left(f(x)\right) = (f_1, f_2, f_3)$ with $\deg(\gcd(f_1, x^n + 1)) = n - k_1, \deg(\gcd(f_2, x^n + 1)) = n - k_2, \deg(\gcd(f_3, x^n + 1)) = n - k_3$. Then, $|C| = 3^{k_1+k_2+k_3}$.

Example 12: Let $C = \langle f(x)\rangle = \langle (2v+2v^2)x^2 + (1+2v+2v^2)x + 1\rangle$ be a negacyclic code of length 3 over R. Hence, $\phi(f(x)) = (x+1, x^2+2x+1, x+1)$ and

$$\begin{aligned} f_1 &= \gcd(x+1, x^3+1) = x+1 \\ f_2 &= \gcd(x^2+2x+1, x^3+1) = x^2+2x+1 \\ f_3 &= \gcd(x+1, x^3+1) = x+1 \end{aligned}$$

So we have $|C| = 3^{2+1+2} = 3^5$.

Example 13: Let $C = \langle f(x)\rangle = \langle v^2x^4 + vx^3 + (1 + 2v^2)x^2 + 2vx + 1\rangle$ be a negacyclic code of length 10 over R. Hence, $\phi(f(x)t) = (x^2+1, x^4+x^3+2x+1, x^4+2x^3+x+1)$ and

$$\begin{aligned} f_1 &= \gcd(x^2+1, x^{10}+1) = x^2+1 \\ f_2 &= \gcd(x^4+x^3+2x+1, x^{10}+1) = x^4+x^3+2x+1 \\ f_3 &= \gcd(x^4+2x^3+x+1, x^{10}+1) = x^4+2x^3+x+1 \end{aligned}$$

So we have $|C| = 3^{8+6+6} = 3^{20}$.

Let $h_i(x) = (x^n+1)/(\gcd(x^n+1, f_i))$. Hence, $C^\perp = \left\langle \phi^{-1}\left(h_{1_R}(x), h_{2_R}(x), h_{3_R}(x)\right)\right\rangle$ where $h_{i_R}(x)$ be the reciprocal polynomial of $h_i(x)$ for $i = 1, 2, 3$. By using the previous Example 13,

$$\begin{aligned} C^\perp &= \left\langle \phi^{-1}\left(h_{1_R}(x), h_{2_R}(x), h_{3_R}(x)\right)\right\rangle \\ &= \langle \phi^{-1}(x^8+2x^6+x^4+2x^2+1, x^6+x^5+x^4+x^2 \\ &\quad +2x+1, x^6+2x^5+x^4+x^2+x+1)\rangle \\ &= \langle (1+2v^2)x^8 + (2+2v^2)x^6 + vx^5 + x^4 \\ &\quad +(2+2v^2)x^2 + 2vx + 1\rangle \end{aligned}$$

IV. Quantum Codes From Cyclic (Negacyclic) Codes Over R

Theorem 14: Let $C_1 = [n, k_1, d_1]_q$ and $C_2 = [n, k_2, d_2]_q$ be linear codes over GF(q) with $C_2^\perp \subseteq C_1$. Furthermore, let $d = min\{wt(v) : v \in (C_1 \backslash C_2^\perp) \cup (C_2^\perp \backslash C_1)\} \geq \min\{d_1, d_2\}$.Then there exists a quantum error-correcting code $C = [n, k_1 + k_2 - n, d]_q$.In particular, if $C_1^\perp \subseteq C_1$, then there exists a quantum error-correcting code $C = [n, n - 2k_1, d_1]$, where $d_1 = min\{wt(v) : v \in (C_1^\perp \backslash C_1)\}$ [16].

Proposition 15: Let $C = (1+2v^2)C_1 \oplus \left(2v + 2v^2\right) C_2 \oplus \left(v + 2v^2\right) C_3$ be a linear code over R. Then C is a cyclic code over R iff C_1, C_2 and C_3 are cyclic codes.

Proof: Let $(a_0, a_1, ..., a_{n-1}) \in C_1$, $(b_0, b_1, ..., b_{n-1}) \in C_2$ and $(c_0, c_1, ..., c_{n-1}) \in C_3$. Assume that $m_i = (1 + 2v^2)a_i + \left(2v + 2v^2\right) b_i + \left(v + 2v^2\right) c_i$ for $i = 0, 1, ..., n - 1$. Then $(m_0, m_1, ..., m_{n-1}) \in C$. Since C is a cyclic code, it follows that $(m_{n-1}, m_0, ..., m_{n-2}) \in C$. Note that $(m_{n-1}, m_0, ..., m_{n-2}) = (1+2v^2)(a_{n-1}, a_0, ..., a_{n-2}) + (2v + 2v^2)(b_{n-1}, b_0, ..., b_{n-2}) + (v+2v^2)(c_{n-1}, c_0, ..., c_{n-2})$. Hence $(a_{n-1}, a_0, ..., a_{n-2}) \in C_1, (b_{n-1}, b_0, ..., b_{n-2}) \in C_2$ and $(c_{n-1}, c_0, ..., c_{n-2}) \in C_3$. Therefore,$C_1, C_2$ and C_3 cyclic codes over Z_3.

Conversely, suppose that C_1, C_2 and C_3 cyclic codes over Z_3. Let $(m_0, m_1, ..., m_{n-1}) \in C$ where $m_i = (1 + 2v^2)a_i + \left(2v + 2v^2\right) b_i + \left(v + 2v^2\right) c_i$ for $i = 0, 1, ..., n - 1$. Then $(a_0, a_1, ..., a_{n-1}) \in C_1, (b_0, b_1, ..., b_{n-1}) \in C_2$ and $(c_0, c_1, ..., c_{n-1}) \in C_3$. Note that $(m_{n-1}, m_0, ..., m_{n-2}) = (1 + 2v^2)(a_{n-1}, a_0, ..., a_{n-2}) + \left(2v + 2v^2\right)(b_{n-1}, b_0, ..., b_{n-2}) + \left(v + 2v^2\right)(c_{n-1}, c_0, ..., c_{n-2}) \in C = (1 + 2v^2)C_1 \oplus \left(2v + 2v^2\right) C_2 \oplus \left(v + 2v^2\right) C_3$. So, C is cyclic code over R. ■

Proposition 16: Let $C = (1+2v^2)C_1 \oplus \left(2v + 2v^2\right) C_2 \oplus \left(v + 2v^2\right) C_3$ be a linear code over R. Then C is a negacyclic code over R iff C_1, C_2 and C_3 are negacyclic codes.

Proof: Let $(a_0, a_1, ..., a_{n-1}) \in C_1$, $(b_0, b_1, ..., b_{n-1}) \in C_2$ and $(c_0, c_1, ..., c_{n-1}) \in C_3$. Assume that $m_i = (1 + 2v^2)a_i + \left(2v + 2v^2\right) b_i + \left(v + 2v^2\right) c_i$ for $i = 0, 1, ..., n-1$. Then $(m_0, m_1, ..., m_{n-1}) \in C$. Since C is a negacyclic code, it follows that $(-m_{n-1}, m_0, ..., m_{n-2}) \in C$. Note that $(-m_{n-1}, m_0, ..., m_{n-2}) = (1 + 2v^2)(-a_{n-1}, a_0, ..., a_{n-2}) + (2v+2v^2)(-b_{n-1}, b_0, ..., b_{n-2}) + (v+2v^2)(-c_{n-1}, c_0, ..., c_{n-2})$. Hence $(-a_{n-1}, a_0, ..., a_{n-2}) \in C_1, (-b_{n-1}, b_0, ..., b_{n-2}) \in C_2$ and $(-c_{n-1}, c_0, ..., c_{n-2}) \in C_3$. Therefore, C_1, C_2 and C_3 negacyclic codes over Z_3.

Conversely, suppose that C_1, C_2 and C_3 negacyclic codes over Z_3. Let $(m_0, m_1, ..., m_{n-1}) \in C$ where $m_i = (1 + 2v^2)a_i + (2v + 2v^2)b_i + (v + 2v^2)c_i$ for $i = 0, 1, ..., n - 1$. Then $(a_0, a_1, ..., a_{n-1}) \in C_1$,

$(b_0, b_1, ..., b_{n-1}) \in C_2$ and $(c_0, c_1, ..., c_{n-1}) \in C_3$. Note that $(-m_{n-1}, m_0, ..., m_{n-2}) = (1+2v^2)(-a_{n-1}, a_0, ..., a_{n-2}) + (2v+2v^2)(-b_{n-1}, b_0, ..., b_{n-2}) + (v+2v^2)(-c_{n-1}, c_0, ..., c_{n-2}) \in C = (1+2v^2)C_1 \oplus (2v+2v^2)C_2 \oplus (v+2v^2)C_3$. So, C is negacyclic code over R. ■

Proposition 17: Suppose $C = (1+2v^2)C_1 \oplus \left(2v+2v^2\right)C_2 \oplus \left(v+2v^2\right)C_3$ is a cyclic (negacyclic) code of length n over R. Then

$$C =< (1+2v^2)f_1, \left(2v+2v^2\right)f_2, \left(v+2v^2\right)f_3 >$$

and $|C| = 3^{3n-(\deg f_1+\deg f_2+\deg f_3)}$ where f_1, f_2 and f_3 generator polynomials of C_1, C_2 and C_3 respectively.

Proposition 18: Suppose C is a cyclic (negacyclic) code of length n over R, then there is a unique polynomial $f\,(x)$ such that $C = \langle f\,(x)\rangle$ and $f\,(x) \mid x^n - 1$ $(f\,(x) \mid x^n + 1)$ where $f\,(x) = (1+2v^2)f_1(x) + \left(2v+2v^2\right)f_2(x) + \left(v+2v^2\right)f_3(x)$.

Proposition 19: Let C be a linear code of length n over R, then $C^\perp = (1+2v^2)C_1^\perp \oplus \left(2v+2v^2\right)C_2^\perp \oplus \left(v+2v^2\right)C_3^\perp$. Furthermore, C is self-dual code iff C_1, C_2 and C_3 are self-dual codes over Z_3.

Proposition 20: If $C = (1+2v^2)C_1 \oplus \left(2v+2v^2\right)C_2 \oplus \left(v+2v^2\right)C_3$ is a cyclic (negacyclic) code of length n over R. Then

$$C^\perp = \left\langle (1+2v^2)h_1^* + \left(2v+2v^2\right)h_2^* + \left(v+2v^2\right)h_3^* \right\rangle$$

and $\left|C^\perp\right| = 3^{\deg f_1+\deg f_2+\deg f_3}$ where for $i = 1,2,3$, h_i^* are the reciprocal polynomials of h_i i.e., $h_i\,(x) = (x^n-1)\,/f_i\,(x)$,$(h_i\,(x) = (x^n+1)\,/f_i\,(x))$, $h_i^*\,(x) = x^{\deg h_i}h_i\left(x^{-1}\right)$ for $i = 1,2,3$.

Lemma 21: A ternary linear cyclic (negacyclic) code C with generator polynomial $f\,(x)$ contains its dual code iff

$$x^n - 1 \equiv 0\,(modff^*)\,, \qquad (x^n + 1 \equiv 0\,(modff^*)\,)$$

where f^* is the reciprocal polynomial of f.

Theorem 22: Let $C = \langle (1+2v^2)f_1, (2v+2v^2)f_2, (v+2v^2)f_3 \rangle$ be a cyclic (negacyclic) code of length n over R. Then $C^\perp \subseteq C$ iff $x^n - 1 \equiv 0\,(modf_if_i^*)$ $(x^n+1 \equiv 0\,(modf_if_i^*))$ for $i = 1,2,3$.

Proof: Let $x^n - 1 \equiv 0\,(modf_if_i^*)$ $(x^n + 1 \equiv 0\,(modf_if_i^*))$ for $i = 1,2,3$. Then $C_1^\perp \subseteq C_1, C_2^\perp \subseteq C_2, C_3^\perp \subseteq C_3$. By using $(1+2v^2)C_1^\perp \subseteq (1+2v^2)C_1$, $\left(2v+2v^2\right)C_2^\perp \subseteq \left(2v+2v^2\right)C_2$, $\left(v+2v^2\right)C_3^\perp \subseteq \left(v+2v^2\right)C_3$. We have $(1+2v^2)C_1^\perp \oplus \left(2v+2v^2\right)C_2^\perp \oplus \left(v+2v^2\right)C_3^\perp \subseteq (1+2v^2)C_1 \oplus \left(2v+2v^2\right)C_2 \oplus \left(v+2v^2\right)C_3$. So, $\langle (1+2v^2)h_1^* + \left(2v+2v^2\right)h_2^* + \left(v+2v^2\right)h_3^* \rangle \subseteq \langle (1+2v^2)f_1, \left(2v+2v^2\right)f_2, \left(v+2v^2\right)f_3 \rangle$. That is $C^\perp \subseteq C$.

Conversely, if $C^\perp \subseteq C$, then $(1+2v^2)C_1^\perp \oplus \left(2v+2v^2\right)C_2^\perp \oplus \left(v+2v^2\right)C_3^\perp \subseteq (1+2v^2)C_1 \oplus \left(2v+2v^2\right)C_2 \oplus \left(v+2v^2\right)C_3$. By thinking $mod(1+2v^2), mod\left(2v+2v^2\right)$ and $mod\left(v+2v^2\right)$ respectively we have $C_i^\perp \subseteq C_i$ for $i = 1,2,3$. Therefore, $x^n - 1 \equiv 0\,(modf_if_i^*)$ $(x^n + 1 \equiv 0\,(modf_if_i^*))$ for $i = 1,2,3$. ■

Corollary 23: $C = (1+2v^2)C_1 \oplus \left(2v+2v^2\right)C_2 \oplus \left(v+2v^2\right)C_3$ is a cyclic (negacyclic) code of length n over R. Then $C^\perp \subseteq C$ iff $C_i^\perp \subseteq C_i$ for $i = 1,2,3$.

Example 24: Let $n = 6, R = Z_3 + vZ_3 + v^2Z_3, v^3 = v$. We have $x^6 - 1 = (2x^2+2)(x^2+2)(2x^2+1) = f_1f_2f_3$ in $Z_3\,[x]$. Hence,

$$\begin{aligned} f_1^* &= 2x^2 + 2 = f_1 \\ f_2^* &= 2x^2 + 1 = f_3 \\ f_3^* &= x^2 + 2 = f_2 \end{aligned}$$

Let $C = \left\langle (1+2v^2)f_2, \left(2v+2v^2\right)f_2, \left(v+2v^2\right)f_3 \right\rangle$. Obviously $x^6 - 1$ is divisibly by $f_if_i^*$ for $i = 2,3$. Thus we have $C^\perp \subseteq C$.

Example 25: Let $n = 10, R = Z_3 + vZ_3 + v^2Z_3, v^3 = v$. We have $x^{10}+1 = (x^2+1)(x^4+x^3+2x+1)(x^4+2x^3+x+1) = g_1g_2g_3$ in $Z_3\,[x]$. Hence,

$$\begin{aligned} g_1^* &= x^2 + 1 = g_1 \\ g_2^* &= x^4 + 2x^3 + x + 1 = g_3 \\ g_3^* &= x^4 + x^3 + 2x + 1 = g_2 \end{aligned}$$

Let $C = \left\langle (1+2v^2)g_2, \left(2v+2v^2\right)g_2, \left(v+2v^2\right)g_3 \right\rangle$. Obviously $x^{10} + 1$ is divisibly by $g_ig_i^*$ for $i = 2,3$. Thus we have $C^\perp \subseteq C$.

Theorem 26: Let C be linear code of length n over R with $|C| = 3^{3k_1+2k_2+k_3}$ and minimum distance d. Then $\phi\,(C)$ is ternary linear $[3n, 3k_1+2k_2+k_3, d]$ code.

Using Theorem 14 and Theorem 22 we can construct quantum codes.

Theorem 27: Let $(1+2v^2)C_1 \oplus \left(2v+2v^2\right)C_2 \oplus \left(v+2v^2\right)C_3$ be a cyclic (negacyclic) code of arbitrary length n over R with type $27^{k_1}9^{k_2}3^{k_3}$. If $C_i^\perp \subseteq C_i$ where $i = 1,2,3$ then $C^\perp \subseteq C$ and there exists a quantum error-correcting code with parameters $[[3n, 2(3k_1+2k_2+k_3) - 3n, d_L]]$ where d_L is the minimum Lee weights of C.

Example 28: Let $n = 6$. We have $x^6 - 1 = (2x^2+2)(x^2+2)(2x^2+1)$ in $Z_3\,[x]$. Let $f_1\,(x) = f_2\,(x) = x^2+2$, $f_3 = 2x^2+1$. Thus $C =< (1+2v^2)f_1, \left(2v+2v^2\right)f_2, \left(v+2v^2\right)f_3 >$. C is a linear cyclic code of length 6.The dual code $C^\perp = \left\langle (1+2v^2)h_1^*, \left(2v+2v^2\right)h_2^*, \left(v+2v^2\right)h_3^* \right\rangle$ can be obtained of Proposition 20. Clearly, $C^\perp \subseteq C$. Hence, we obtain a quantum code with parameters $[[18, 6, 2]]$.

Example 29: Let $n = 8$. We have $x^8 - 1 = (x+1)(x+2)(x^2+1)(x^2+x+2)(x^2+2x+2)$ in $Z_3\,[x]$. Let $f_1\,(x) = f_2\,(x) = f_3\,(x) = x^2+1$. Thus $C = \langle (1+2v^2)f_1, \left(2v+2v^2\right)f_2, \left(v+2v^2\right)f_3 \rangle$. C is a linear cyclic code of length 8. Hence, we obtain a quantum code with parameters $[[24, 12, 2]]$.

Example 30: Let $n = 12$. We have $x^{12} - 1 = (x-1)^3\left(x^3+x^2+x+1\right)^3$ in $Z_3\,[x]$. Let $f_1\,(x) = f_2\,(x) = f_3\,(x) = x^3+x^2+x+1$.Thus $C = \langle (1+2v^2)f_1, \left(2v+2v^2\right)f_2, \left(v+2v^2\right)f_3 \rangle$. C is a linear cyclic code of length 12. The dual code $C^\perp = \langle (1+2v^2)h_1^*, (2v+2v^2)h_2^*, (v+2v^2)h_3^* \rangle$ can be obtained of Proposition 20. Clearly, $C^\perp \subseteq C$. Hence, we obtain a quantum code with parameters $[[36, 18, 2]]$.

Let $n = 27$. We have $x^{27} - 1 = (x-1)^3(x^3-1)^4(x^6-2x^3+1)^2$ in $Z_3[x]$. Let $f_1(x) = f_2(x) = f_3(x) = x^6-2x^3+1$. Hence, we obtain a quantum code with parameters $[[81, 45, 2]]$.

Let $n = 30$. We have $x^{30} - 1 = (x^2 + 2)^3(x^4 + x^3 + x^2 + x + 1)^3(x^4 + 2x^3 + x^2 + 2x + 1)^3$ in $Z_3[x]$. Let $f_1(x) = f_3(x) = x^4 + x^3 + x^2 + x + 1$, $f_2(x) = x^4 + 2x^3 + x^2 + 2x + 1$. Hence, we obtain a quantum code with parameters $[[90, 66, 2]]$.

Example 31: Let $n = 3$. We have $x^3 + 1 = (x + 1)^3$ in $Z_3[x]$. Let $f_1(x) = f_2(x) = f_3(x) = x + 1$. Thus $C = \left\langle (1 + 2v^2)f_1, \left(2v + 2v^2\right) f_2, \left(v + 2v^2\right) f_3 \right\rangle$. C is a linear negacyclic code of length 3. The dual code $C^\perp =< (1 + 2v^2)h_1^*, (2v + 2v^2)h_2^*, \left(v + 2v^2\right) h_3^* >$ can be obtained of Proposition 20. Clearly, $C^\perp \subseteq C$. Hence, we obtain a quantum code with parameters $[[9, 3, 2]]$.

Example 32: Let $n = 10$. We have $x^{10} + 1 = (x^2 + 1)(x^4 + x^3 + 2x + 1)(x^4 + 2x^3 + x + 1)$ in $Z_3[x]$. Let $f_1(x) = x^4 + x^3 + 2x + 1$, $f_2(x) = f_3(x) = x^4 + 2x^3 + x + 1$. Thus $C = \left\langle (1 + 2v^2)f_1, \left(2v + 2v^2\right) f_2, \left(v + 2v^2\right) f_3 \right\rangle$. C is a linear negacyclic code of length 10. The dual code $C^\perp = \left\langle (1 + 2v^2)h_1^*, \left(2v + 2v^2\right) h_2^*, \left(v + 2v^2\right) h_3^* \right\rangle$ can be obtained of Proposition 20. Clearly, $C^\perp \subseteq C$. Hence, we obtain a quantum code with parameters $[[30, 6, 4]]$.

Example 33: Let $n = 12$. We have $x^{12} + 1 = (x^4 + 1)(x^2 + x + 2)(x^2 + 2x + 2)(2x^2 + 2x + 1 \left(2x^2 + x + 1\right)$ in $Z_3[x]$. Let $f_1(x) = x^2 + x + 2$, $f_2(x) = 2x^2 + x + 1$, $f_3(x) = x^2 + 2x + 2$. Thus $C = \left\langle (1 + 2v^2)f_1, \left(2v + 2v^2\right) f_2, \left(v + 2v^2\right) f_3 \right\rangle$. C is a linear negacyclic code of length 12. The dual code $C^\perp =< (1 + 2v^2)h_1^*, \left(2v + 2v^2\right) h_2^*, \left(v + 2v^2\right) h_3^* >$ can be obtained of Proposition 20. Clearly, $C^\perp \subseteq C$. Hence, we obtain a quantum code with parameters $[[36, 24, 2]]$.

V. Constacyclic codes over R

Let $\lambda = \alpha + \beta v + \gamma v^2$ be unit element of R. Note that $\lambda^n = 1$ if n even $\lambda^n = \lambda$ if n odd. So we only study λ-constacyclic codes of odd length.

Proposition 34: Let ϱ be the map of $R[x] / \langle x^n - 1 \rangle$ into $R[x] / \langle x^n - \lambda \rangle$ defined by $\varrho(a(x)) = a(\lambda x)$. If n is odd, then ϱ is a ring isomorphism.

Proof: The proof is straightforward if n is odd, $a(x) \equiv b(x)(mod(x^n - 1))$ iff $a(\lambda x) \equiv b(\lambda x)(mod(x^n - \lambda))$ ■

Corollary 35: I is an ideal of $R[x] / \langle x^n - 1 \rangle$ if and only if $\varrho(I)$ is an ideal of $R[x] / \langle x^n - \lambda \rangle$.

Corollary 36: Let $\overline{\varrho}$ be the permutation of R^n with n odd, such that $\overline{\varrho}(a_0, a_1, ..., a_{n-1}) = (a_0, \lambda a_1, \lambda^2 a_2 ..., \lambda^{n-1} a_{n-1})$ and C be a subset of R^n then C is a linear cyclic code iff $\overline{\varrho}(C)$ is a linear λ-constacyclic code.

Corollary 37: C is a cyclic code of parameters $(n, 3^k, d)$ over R iff $\overline{\varrho}(C)$ is a λ-constacyclic code of parameters $(n, 3^k, d)$ over R, when n is odd.

Theorem 38: Let λ be a unit in R. Let $C = (1 + 2v^2)C_1 \oplus \left(2v + 2v^2\right) C_2 \oplus \left(v + 2v^2\right) C_3$ be a linear code of length n over R. Then C is a λ-constacyclic code of length n over R iff C_i are either cyclic codes or negacyclic codes of length n over Z_3 for $i = 1, 2, 3$.

Proof: Let ν be the λ-constacyclic shift on R^n. Let C be a λ-constacyclic code of length n over R.Let $(a_0, a_1, ..., a_{n-1}) \in C_1, (b_0, b_1, ..., b_{n-1}) \in C_2$ and $(c_0, c_1, ..., c_{n-1}) \in C_3$. Then the corresponding element of C is $(m_0, m_1, ..., m_{n-1}) = (1 + 2v^2)(a_0, a_1, ..., a_{n-1}) + (2v + 2v^2)(b_0, b_1, ..., b_{n-1}) + (v + 2v^2)(c_0, c_1, ..., c_{n-1})$. Since C is a λ-constacyclic code so, $\nu(m) = (\lambda m_{n-1}, m_0, ..., m_{n-2}) \in C$ where $m_i = a_i + b_i v + v^2 c_i$ for $i = 0, 1, ..., n-1$. Let $\lambda = \alpha + v\beta + v^2\gamma$, where $\alpha, \beta, \gamma \in Z_3$. $\nu(m) = (1 + 2v^2)(\lambda a_{n-1}, a_0, ..., a_{n-2}) + (2v + 2v^2)(\lambda b_{n-1}, b_0, ..., b_{n-2}) + (v + 2v^2)(\lambda c_{n-1}, c_0, ..., c_{n-2})$. Since the units of Z_3 are 1 and -1, so $\alpha = \mp 1$. Therefore we have obtained the desired result. The other side it is seen easily. ■

VI. Skew Codes Over R

We are interested in studying skew codes using the ring $R = Z_3 + vZ_3 + v^2 Z_3$ where $v^3 = v$. We define non-trivial ring automorphism θ on the ring R by $\theta\left(a + vb + v^2 c\right) = a + 2bv + v^2 c$ for all $a + vb + v^2 c \in R$.

The ring $R[x, \theta] = \{a_0 + a_1 x + ... + a_{n-1}x^{n-1} : a_i \in R, n \in N\}$ is called a skew polynomial ring. This ring is a non-commutative ring. The addition in the ring $R[x, \theta]$ is the usual polynomial addition and multiplication is defined using the rule, $(ax^i)(bx^j) = a\theta^i(b)x^{i+j}$. Note that $\theta^2(a) = a$ for all $a \in R$. This implies that θ is a ring automorphism of order 2.

Definition 39: A subset C of R^n is callled a skew cyclic code of length n if C satisfies the following conditions,
$i)$ C is a submodule of R^n,
$ii)$ If $c = (c_0, c_1, ..., c_{n-1}) \in C$, then $\sigma_\theta(c) = \left(\theta(c_{n-1}), \theta(c_0), ..., \theta(c_{n-2})\right) \in C$.

Let $f(x) + (x^n - 1)$ be an element in the set $R_n = R[x, \theta] / (x^n - 1)$ and let $r(x) \in R[x, \theta]$. Define multiplication from left as follows,

$$r(x)(f(x) + (x^n - 1)) = r(x)f(x) + (x^n - 1)$$

for any $r(x) \in R[x, \theta]$.

Theorem 40: R_n is a left $R[x, \theta]$-module where multiplication defined as in above.

Theorem 41: A code C in R_n is a skew cyclic code if and only if C is a left $R[x, \theta]$-submodule of the left $R[x, \theta]$-module R_n.

Theorem 42: Let C be a skew cyclic code in R_n and let $f(x)$ be a polynomial in C of minimal degree. If $f(x)$ is monic polynomial, then $C = (f(x))$ where $f(x)$ is a right divisor of $x^n - 1$.

Theorem 43: A module skew cyclic code of length n over R is free iff it is generated by a monic right divisor $f(x)$ of $x^n - 1$. Moreover, the set $\{f(x), xf(x), x^2 f(x), ..., x^{n - \deg(f(x)) - 1} f(x)\}$ forms a basis of C and the rank of C is $n - \deg(f(x))$.

Theorem 44: Let n be odd and C be a skew cyclic code of length n. Then C is equivalent to cyclic code of length n over R.

Proof: Since n is odd, $gcd(2, n) = 1$. Hence there exist integers b, c such that $2b + nc = 1$. So $2b = 1 - nc = 1 + zn$ where $z > 0$. Let $a(x) = a_0 + a_1 x + ... + a_{n-1}x^{n-1}$ be a codeword in C. Note that $x^{2b}a(x) = \theta^{2b}(a_0)x^{1+zn} +$

$\theta^{2b}(a_1)x^{2+zn} + ... + \theta^{2b}(a_{n-1})x^{n+zn} = a_{n-1} + a_0x + ... + a_{n-2}x^{n-2} \in C$. Thus C is a cyclic code of length n. ■

Corollary 45: Let n be odd. Then the number of distinct skew cyclic codes of length n over R is equal to the number of ideals in $R\left[x\right]/(x^n-1)$ because of Theorem 44. If $x^n - 1 = \prod_{i=0}^{r} p_i^{s_i}(x)$ where $p_i(x)$ are irreducible polynomials over Z_3. Then the number of distinct skew cyclic codes of length n over R is $\prod_{i=0}^{r}(s_i+1)^3$.

Example 46: Let $n = 27$ and $f(x) = x^3 - 1$. Then $f(x)$ generates a skew cyclic codes of length 27. This code is equivalent to a cyclic code of length 27. Since $x^{27} - 1 = (x-1)^3(x^3-1)^4(x^6-2x^3+1)^2$, it follows that there are 60^3 skew cyclic code of length 27.

Definition 47: A subset C of R^n is called a skew quasi-cyclic code of length n if C satisfies the following conditions,
$i)$ C is a submodule of R^n,
$ii)$ If $e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, .., e_{s-1,l-1}) \in C$, then
$\tau_{\theta,s,l}\left(e\right) = (\theta(e_{s-1,0}), ..., \theta(e_{s-1,l-1}), \theta(e_{0,0}), ..., \theta(e_{0,l-1}), ..., \theta(e_{s-2,0}), ..., \theta(e_{s-2,l-1})) \in C$.

We note that $x^s - 1$ is a two sided ideal in $R\left[x, \theta\right]$ if $m|s$ where m is the order of θ and equal to two. So $R\left[x,\theta\right]/(x^s - 1)$ is well defined.

The ring $R_s^l = (R\left[x,\theta\right]/(x^s-1))^l$ is a left $R_s = R\left[x,\theta\right]/(x^s-1)$ module by the following multiplication on the left $f(x)(g_1(x), ..., g_l(x)) = (f(x)g_1(x), ...f(x)g_l(x))$. If the map γ is defined by

$$\gamma : R^n \longrightarrow R_s^l$$

$(e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1}) \mapsto (e_0(x), ..., e_{l-1}(x))$ such that $e_j(x) = \sum_{i=0}^{s-1} e_{i,j}x^i \in R_s^l$ where $j = 0, 1, ..., l-1$ then the map γ gives a one to one correspondence R^n and the ring R_s^l.

Theorem 48: A subset C of R^n is a skew quasi-cyclic code of length $n = sl$ and index l if and only if $\gamma(C)$ is a left R_s-submodule of R_s^l.

A code C is said to be skew constacyclic if C is closed the under the skew constacyclic shift $\sigma_{\theta,\lambda}$ from R^n to R^n defined by $\sigma_{\theta,\lambda}\left((c_0, c_1, ..., c_{n-1})\right) = \left(\theta\left(\lambda c_{n-1}\right), \theta\left(c_0\right), ..., \theta\left(c_{n-2}\right)\right)$.

Privately, such codes are called skew cyclic and skew negacyclic codes when λ is 1 and -1, respectively.

Theorem 49: A code C of length n over R is skew constacyclic iff the skew polynomial representation of C is a left ideal in $R\left[x,\theta\right]/(x^n - \lambda)$.

VII. The Gray Images of Skew Codes Over R

Proposition 50: Let σ_θ be the skew cyclic shift on R^n, let ϕ be the Gray map from R^n to Z_3^{3n} and let φ be as in the preliminaries. Then $\phi\sigma_\theta = \rho\varphi\phi$ where $\rho(x,y,z) = (x,z,y)$ for every $x, y, z \in Z_3^n$.

Proof: Let $r_i = a_i + vb_i + v^2c_i$ be the elements of R, for $i = 0, 1, ..., n-1$. We have $\sigma_\theta\left(r_0, r_1, ..., r_{n-1}\right) = (\theta(r_{n-1}), \theta(r_0), ..., \theta(r_{n-2}))$. If we apply ϕ, we have

$$\begin{aligned}\phi\left(\sigma_\theta\left(r_0, ..., r_{n-1}\right)\right) &= \phi(\theta(r_{n-1}), \theta(r_0), ..., \theta(r_{n-2}))\\ &= (a_{n-1}, ..., a_{n-2}, a_{n-1} + 2b_{n-1} +\\ &\quad c_{n-1}, ..., a_{n-2} + 2b_{n-2} + c_{n-2},\\ &\quad a_{n-1} + b_{n-1} + c_{n-1}, ..., a_{n-2} +\\ &\quad b_{n-2} + c_{n-2})\end{aligned}$$

On the other hand, $\phi(r_0, ..., r_{n-1}) = (a_0, ..., a_{n-1}, a_0 + b_0 + c_0, ..., a_{n-1} + b_{n-1} + c_{n-1}, a_0 + 2b_0 + c_0, ..., a_{n-1} + 2b_{n-1} + c_{n-1})$. If we apply φ, we have

$\varphi\left(\phi\left(r_0, r_1, ..., r_{n-1}\right)\right) = (a_{n-1}, ..., a_{n-2}, a_{n-1} + b_{n-1} + c_{n-1}, ..., a_{n-2} + b_{n-2} + c_{n-2}, a_{n-1} + 2b_{n-1} + c_{n-1}, ..., a_{n-2} + 2b_{n-2} + c_{n-2})$. If we apply ρ, we have $\rho(\varphi\left(\phi\left(r_0, r_1, ..., r_{n-1}\right)\right)) = (a_{n-1}, ..., a_{n-2}, a_{n-1} + 2b_{n-1} + c_{n-1}, ..., a_{n-2} + 2b_{n-2} + c_{n-2}, a_{n-1} + b_{n-1} + c_{n-1}, ..., a_{n-2} + b_{n-2} + c_{n-2})$. So, we have $\phi\sigma_\theta = \rho\varphi\phi$. ■

Theorem 51: The Gray image a skew cyclic code over R of length n is permutation equivalent to quasi-cyclic code of index 3 over Z_3 with length $3n$.

Proof: Let C be a skew cyclic codes over S of length n. That is $\sigma_\theta(C) = C$. If we apply ϕ, we have $\phi(\sigma_\theta(C)) = \phi(C)$. From the Proposition 50, $\phi(\sigma_\theta(C)) = \phi(C) = \rho(\varphi(\phi(C)))$. So, $\phi(C)$ is permutation equivalent to quasi-cyclic code of index 3 over Z_3 with length $3n$. ■

Proposition 52: Let $\tau_{\theta,s,l}$ be skew quasi-cyclic shift on R^n, let ϕ be the Gray map from R^n to Z_3^{3n}, let Γ be as in the preliminaries, let ρ be as above. Then $\phi\tau_{\theta,s,l} = \rho\Gamma\phi$.

Theorem 53: The Gray image a skew quasi-cyclic code over R of length n with index l is permutation equivalent to l quasi-cyclic code of index 3 over Z_3 with length $3n$.

Proposition 54: Let $\sigma_{\theta,\lambda}$ be skew constacyclic shift on R^n, let ϕ be the Gray map from R^n to Z_3^{3n}, let ρ be as above. Then $\phi\nu = \rho\phi\sigma_{\theta,\lambda}$.

Theorem 55: The Gray image a skew constacyclic code over R of length n is permutation equivalent to the Gray image of a constacyclic code over Z_3 with length $3n$.

The proofs of Proposition 52, 54 and Theorem 53, 55 are similiar to the proofs Proposition 50 and Theorem 51.

VIII. Quasi-constacyclic and Skew Quasi-constacyclic Codes over R

Let $M_s = R\left[x\right]/\left\langle x^s - \lambda\right\rangle$ where λ is a unit element of R.

Definition 56: A subset C of R^n is a called a quasi-constacyclic code of length $n = ls$ with index l if
$i)$ C is a submodule of R^n,
$ii)$ if $e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1}) \in C$ then
$\nabla_{\lambda,l}\left(e\right) = (\lambda e_{s-1,0}, ..., \lambda e_{s-1,l-1}, e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-2,0}, ..., e_{s-2,l-1}) \in C$.

When $\lambda = 1$ the quasi-constacyclic codes are just quasi-cyclic codes.

Since $x^s - \lambda = f_1(x)f_2(x)...f_r(x)$, it follows that

$(R[x]/(x^s-\lambda))^l \cong (R[x]/(f_1(x)))^l \times (R[x]/(f_2(x)))^l \times ... \times (R[x]/(f_r(x)))^l$.

Every submodule of $(R[x]/(x^s-\lambda))^l$ is a direct product of submodules of $(R[x]/(f_t(x)))^l$ for $1 \le t \le r$.

Theorem 57: If $(s,3)=1$ then a quasi-constacyclic code of length $n=sl$ with index l over R is a direct product of linear codes over $R[x]/(f_t(x))$ for $1 \le t \le r$.

Let $x^s-\lambda = f_1(x)f_2(x)...f_r(x)$ be the factorization of $x^s-\lambda$ into irreducible polynomials. Thus, if $(s,3)=1$ and C_i is a linear code of length l over $R[x]/(f_t(x))$ for $1 \le t \le r$, then $\prod_{t=1}^{r} C_t$ is a quasi-constacyclic code of length $n=sl$ over R with $\prod_{t=1}^{r} |C_t|$ codewords.

Define a map $\chi : R^n \to M_s^l$ by $\chi(e) = (e_0(x), e_1(x), ..., e_{l-1}(x))$ where $e_j(x) = \sum_{i=o}^{s-1} e_{ij}x^i \in M_s$, $j=0,1,...,l-1$.

Lemma 58: Let $\chi(C)$ denote the image of C under χ. The map χ induces a one to one correspondence between quasi-constacyclic codes over R of length n with index l and linear codes over M_s of length l.

We define a conjugation map on M_s as one that acts as the identity on the elements of R and that sends x to $x^{-1}=x^{s-1}$, and extended linearly.

We define on $R^{n=sl}$ the usual Euclidean inner product for

$$e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1})$$

and

$$c = (c_{0,0}, ..., c_{0,l-1}, c_{1,0}, ..., c_{1,l-1}, ..., c_{s-1,0}, ..., c_{s-1,l-1})$$

we define $e.c = \sum_{i=0}^{s-1}\sum_{j=0}^{l-1} e_{ij}c_{ij}$.

On M_s^l, we define the Hermitian inner product for $a(x) = (a_0(x), a_1(x), ..., a_{l-1}(x))$ and $b(x) = (b_0(x), b_1(x), ..., b_{l-1}(x))$,

$$\langle a,b \rangle = \sum_{j=0}^{l-1} a_j(x)\overline{b_j(x)}.$$

Theorem 59: Let $e,c \in R^n$. Then $\left(\nabla_{\lambda,l}^k(e)\right).c = 0$ for all $k=0,...,s-1$ iff $\langle \chi(e), \chi(c) \rangle = 0$.

Corollary 60: Let C be a quasi-constacyclic code of length sl with index l over R and $\chi(C)$ be its image in M_s^l under χ. Then $\chi(C)^{\perp} = \chi(C^{\perp})$, where the dual in R^{sl} is taken with respect to the Euclidean inner product, while the dual in M_s^l is taken with respect to the Hermitian inner product. The dual of a quasi-constacyclic code of length sl with index l over R is a quasi-constacyclic code of length sl with index l over .

From [22] we get the following results.

Theorem 61: Let C be a quasi-constacyclic code of length n=sl with index l over R. Let $C^{\perp}$ is the dual of C. If $C = C_1 \oplus C_2 \oplus ... \oplus C_r$ then $C^{\perp} = C_1^{\perp} \oplus C_2^{\perp} \oplus ... \oplus C_r^{\perp}$.

Theorem 62: Let $C = C_1 \oplus C_2 \oplus ... \oplus C_r$ be a quasi-constacyclic code of length n=sl with index l over R where C_t is a free linear code of length l with rank k_t over $R[x]/(f_t(x))$for $1 \le t \le r$. Then C is a κ-generator quasi-constacyclic code and $C^{\perp}$ is an $(l-\kappa')$-generator quasi-constacyclic code where $\kappa = \max_t(k_t)$ and $\kappa' = \min_t(k_t)$.

Let $M_{\theta,s} = R[x,\theta]/\langle x^s-\lambda \rangle$ where λ is a unit element of R. Let θ be an automorphism of R with $|\langle\theta\rangle| = m = 2$.

Definition 63: A subset C of R^n is a called a skew quasi-constacyclic code of length $n=ls, m|s$, with index l if
i) C is a submodule of R^n,
ii) if $e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1})$ $\in C$ then
$\nabla_{\theta,\lambda,l}(e) = (\theta(\lambda e_{s-1,0}), ..., \theta(\lambda e_{s-1,l-1}), \theta(e_{0,0}), ..., \theta(e_{0,l-1})$ $, \theta(e_{1,0}), ..., \theta(e_{1,l-1}), ..., \theta(e_{s-2,0}), ..., \theta(e_{s-2,l-1})) \in C$.

When $\lambda = 1$ the skew quasi-constacyclic codes are just skew quasi-cyclic codes.

The ring $M_{\theta,s}^l$ is a left $M_{\theta,s}$ module where we define multiplication from left by $f(x)(g_1(x), ..., g_l(x)) = (f(x)g_1(x), ...f(x)g_l(x))$.

Define a map $\Lambda : R^n \to M_{\theta,s}^l$ by $\Lambda(e) = (e_0(x), e_1(x), ..., e_{l-1}(x))$ where $e_j(x) = \sum_{i=o}^{s-1} e_{ij}x^i \in M_{\theta,s}$, $j=0,1,...,l-1$.

Lemma 64: Let $\Lambda(C)$ denote the image of C under Λ. The map Λ induces a one to one correspondence between skew quasi-constacyclic codes over R of length n with index l and linear codes over $M_{\theta,s}$ of length l.

Theorem 65: A subset C of R^n is a skew quasi-constacyclic code of length $n=ls$ with index l iff is a left submodule of the ring $M_{\theta,s}^l$.

Proof: Let C be a skew quasi-constacyclic code of index l over R.Suppose that $\Lambda(C)$ forms a submodule of $M_{\theta,s}^l$. $\Lambda(C)$ is closed under addition and scalar multiplication. Let $\Lambda(e) = (e_0(x), e_1(x), ..., e_{l-1}(x)) \in \Lambda(C)$ for $e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1}) \in C$. Then $x\Lambda(e) \in \Lambda(C)$. By linearity it follows that $r(x)\Lambda(e) \in \Lambda(C)$ for any $r(x) \in M_{\theta,s}$. Therefore, $\Lambda(C)$ is a left module of $M_{\theta,s}^l$.

Conversely, suppose E is an $M_{\theta,s}$ left submodule of $M_{\theta,s}^l$. Let $C = \Lambda^{-1}(E) = \{e \in R^n : \Lambda(e) \in E\}$. We claim that C is a skew quasi-constacyclic code of R. Since Λ is a isomorphism, C is a linear code of length n over R. Let $e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1}) \in C$. Then $\Lambda(e) = (e_0(x), e_1(x), ..., e_{l-1}(x)) \in \Lambda(C)$, where $e_j(x) = \sum_{i=o}^{s-1} e_{ij}x^i \in M_{\theta,s}$ for $j=0,1,...,l-1$. It is easy to see that $\Lambda(\nabla_{\theta,\lambda,l}(e)) = x(e_0(x), e_1(x), ..., e_{l-1}(x)) = (xe_0(x), xe_1(x), ..., xe_{l-1}(x)) \in E$. Hence $\nabla_{\theta,\lambda,l}(e) \in C$. So, C is a skew quasi-constacyclic code C. ■

On $R^{n=sl}$ the usual Euclidean inner product for

$$e = (e_{0,0}, ..., e_{0,l-1}, e_{1,0}, ..., e_{1,l-1}, ..., e_{s-1,0}, ..., e_{s-1,l-1})$$

and

$$c = (c_{0,0}, ..., c_{0,l-1}, c_{1,0}, ..., c_{1,l-1}, ..., c_{s-1,0}, ..., c_{s-1,l-1})$$

we define $e.c = \sum_{i=0}^{s-1}\sum_{j=0}^{l-1} e_{ij}c_{ij}$. We define a conjugation map Ω on $M_{\theta,s}^l$ such that $\Omega(cx^i) = \theta^{-1}(c)x^{s-1}, 0 \le i \le s-1$, and extended linearly. We define the Hermitian inner product for $a = (a_0(x), a_1(x), ..., a_{l-1}(x))$ and $b = (b_0(x), b_1(x), ..., b_{l-1}(x))$,

$$\langle a, b\rangle = \sum_{j=0}^{l-1} a_j(x)\Omega\left(b_j(x)\right).$$

Theorem 66: Let $e, c \in R^n$. Then $\left(\nabla^k_{\theta,\lambda,l}(e)\right).c = 0$ for all $k = 0, ..., s-1$ iff $\langle \Lambda(e), \Lambda(c)\rangle = 0$.

Proof: Since $\theta^s = 1, \langle e, c\rangle = 0$ is equivalent to

$$\begin{aligned} 0 &= \sum_{j=0}^{l-1} e_j(x)\Omega\left(c_j(x)\right) = \sum_{j=0}^{l-1}\left(\sum_{i=0}^{s-1} e_{ij}x^i\right)\Omega\left(\sum_{k=0}^{s-1} c_{kj}x^k\right) \\ &= \sum_{j=0}^{l-1}\left(\sum_{i=0}^{s-1} e_{ij}x^i\right)\left(\sum_{k=0}^{s-1}\theta^{-1}(c_{kj})x^{s-k}\right) \\ &= \sum_{j=0}^{l-1}\left(\sum_{j=0}^{l-1}\sum_{i=0}^{s-1} e_{i+h,j}\theta^h(c_{ij})\right)x^h \end{aligned}$$

where the subscript $i+h$ is taken modulo s. Equating the coefficients of x^h on both sides, we have $\sum_{j=0}^{l-1}\sum_{i=0}^{s-1} w_{i+h,j}\theta^h(c_{ij}) = 0$, for all $0 \leq h \leq s-1$. $\sum_{j=0}^{l-1}\sum_{i=0}^{s-1} e_{i+h,j}\theta^h(c_{ij}) = 0$ is equivalent to $\theta^h(\nabla^{s-h}_{\theta,\lambda,l}(e).c) = 0$ which is further equivalent to $\nabla^{s-h}_{\theta,\lambda,l}(e,c) = 0$, for all $0 \leq h \leq s-1$. Since $0 \leq h \leq s-1$, condition is equivalent to $\left(\nabla^k_{\theta,\lambda,l}(e)\right).c = 0$ for all $k = 0, ..., s-1$. ■

Corollary 67: Let C be a skew quasi-constacyclic code of length $n = sl$ with index l over R. Then $C^\perp = \left\{a(x) \in M^l_{\theta,s} \ : \ \langle a(x), b(x)\rangle = 0, \ \forall\ b(x) \in C\right\}$.

Corollary 68: Let C be a skew quasi-constacyclic code of length sl with index l over R and $\Lambda(C)$ be its image in $M^l_{\theta,s}$ under Λ. Then $\Lambda(C)^\perp = \Lambda\left(C^\perp\right)$, where the dual in R^{sl} is taken with respect to the Euclidean inner product, while the dual in $M^l_{\theta,s}$ is taken with respect to the Hermitian inner product. The dual of a skew quasi-constacyclic code of length sl with index l over R is a skew quasi-constacyclic code of length sl with index l over R.

Proposition 69: Let $\nabla_{\theta,\lambda,l}$ be skew quasi-constacyclic shift on R^n, let ϕ be the Gray map from R^n to Z_3^{3n}. Then $\phi\nabla_{\lambda,l} = \rho\phi\nabla_{\theta,\lambda,l}$, where $\rho(x, y, z) = (x, z, y)$ for every $x, y, z \in Z_3^n$.

Proof: The proof is similar to the proof of Proposition 50. ■

Theorem 70: The Gray image a skew quasi-constacyclic code over R of length n is permutation equivalent to the Gray image of a quasi-constacyclic code over Z_3 with length $3n$.

Proof: The proof is similar to the proof of Theorem 51. ■

IX. 1-generator Skew Quasi-constacyclic Codes over R

A 1-generator skew quasi-constacyclic code over R is a left $M_{\theta,s}$-submodule of $M^l_{\theta,s}$ generated by $\mathbf{f(x)} = (f_1(x), f_2(x), ..., f_l(x)) \in M^l_{\theta,s}$ has the form $C = \{g(x)\left(f_1(x), f_2(x), ..., f_l(x)\right) : \ g(x) \in M_{\theta,s}\}$. Define the following map

$$\Pi_i : M^l_{\theta,s} \longrightarrow M_{\theta,s}$$

defined by $(e_1(x), e_2(x), ..., e_l(x)) \longmapsto e_i(x)$, $1 \leq i \leq l$. Let $\Pi_i(C) = C_i$. Since C is a left $M_{\theta,s}$-submodule of $M^l_{\theta,s}$, C_i is a left $M_{\theta,s}$-submodule of $M_{\theta,s}$,that is a left ideal of $M_{\theta,s}$. C_i is generated by $f_i(x)$. Hence C_i is a principal skew constacyclic code of length n over R. $f_i(x)$ is a monic right divisor of $x^s - \lambda$ that is $x^s - \lambda = h_i(x)f_i(x)$, $1 \leq i \leq l$.

A generator of C has the form

$$\mathbf{f(x)} = (g_1(x)f_1(x), g_2(x)f_2(x), ..., g_l(x)f_l(x))$$

where $g_i(x) \in R[x, \theta]$ such that $g_i(x)$ and $h_i(x)$ are right coprime for all $1 \leq i \leq l$.

Definition 71: Let $C = (g_1(x)f_1(x), g_2(x)f_2(x), ..., g_l(x)$ $f_l(x))$ be a skew quasi-constacyclic code of length $n = sl$ with index l. Then unique monic polynomial

$$f(x) = gcld(\mathbf{f(x)}, x^s - \lambda) = gcld(f_1(x), f_2(x), ..., f_l(x), x^s - \lambda)$$

is called the generator polynomial of C.

Theorem 72: Let C be a 1-generator skew quasi-constacyclic code of length $n = sl$ with index l over R generated by $\mathbf{f(x)} = (f_1(x), f_2(x), ..., f_l(x))$ where $f_i(x)$ is a monic divisor of $x^s - \lambda$. Then C is a R-free code with rank $s - deg(f(x))$ where $f(x) = gcld(\mathbf{f(x)}, x^s - \lambda)$. Moreover, the set $\{\mathbf{f(x)}, x\mathbf{f(x)}, ..., x^{n-\deg(f(x))-1}\mathbf{f(x)}\}$ forms an R-basis of C.

Proof: Since $gcld(f_i(x), x^s - \lambda) = m_i(x)$, it follows that $f(x) = gcld(m_1(x), m_2(x), ..., m_l(x))$ where $\Pi_i(C) = (f_i(x)) = (m_i(x))$ with $m_i(x)|(x^s - \lambda)$ for all $1 \leq i \leq l$. Let $c(x) = \sum_{i=0}^{n-k-1} c_i x^i$ and $c(x)\mathbf{f(x)} = 0$. Then $(x^s - \lambda)|c(x)f_i(x)$ for all $1 \leq i \leq l$. Hence $(x^s - \lambda)|c(x)f_i(x)c_i(x)$ with $gcld(c_i(x), \frac{x^s-\lambda}{f_i(x)}) = 1$. That is $\frac{x^s-\lambda}{f_i(x)}|c(x)$ which implies that $\frac{x^s-\lambda}{f(x)}|c(x)$. Since $\deg(\frac{x^s-\lambda}{f(x)}) = s - k > \deg(c(x)) = n - k - 1$, it is follows that $c(x) = 0$. Thus, $\mathbf{f(x)}, x\mathbf{f(x)}, ..., x^{n-\deg(f(x))-1}\mathbf{f(x)}$ are R-linear independent. Further, $\mathbf{f(x)}, x\mathbf{f(x)}, ..., x^{n-\deg(f(x))-1}\mathbf{f(x)}$ generate C. So, $\left\{\mathbf{f(x)}, x\mathbf{f(x)}, ..., x^{n-\deg(f(x))-1}\mathbf{f(x)}\right\}$ forms an R-basis of C. ■

Conclusion

In this paper, we have introduced skew cyclic, skew quasi-cyclic, skew constacyclic and skew quasi-constacyclic codes over the finite ring R. By using the Gray map, we have studied the Gray images of cyclic, quasi-cyclic, constacyclic and their skew codes over R. We have obtained a representation of a linear code of length n over R using C_1, C_2 and C_3 which are linear codes of length n over Z_3. We have obtained the parameters of quantum error-correcting codes from both cyclic and negacyclic codes over R. We have determined a sufficient condition for 1-generator skew quasi-constacyclic codes to be free.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their valuable remarks and suggestions. We would like to thank S.T. Dougherty for helping to express a linear code C of length n over R by means of three ternary codes and we would like to thank Sedat Akleylek for helping to find minimum distances by using computer program.

REFERENCES

[1] T. Abualrub, A. Ghrayeb, N. Aydın, I. Siap, *On the construction of skew quasi-cyclic codes*, IEEE Transsactions on Information Theory, **56** 2081-2090, (2010).

[2] T. Abualrub, N. Aydın, P. Seneviratne, *On θ-cyclic codes over F_2+vF_2*, Australasian Journal of Combinatorics, **54** 115-126, (2012).

[3] M. Ashraf, G. Mohammad, *Quantum codes from cyclic codes over F_3+vF_3*, International Journal of Quantum Information, **6** 1450042, (2014).

[4] A. Bayram, I. Siap, *Structure of codes over the ring $Z_3[v]/<v^3-v>$*, AAECC, DOI 10.1007/s00200-013-0208-x, (2013).

[5] M. Bhaintwal, *Skew quasi-cyclic codes over Galois rings*, Des. Codes Cryptogr., DOI 10.1007/s10623-011-9494-0.

[6] M. Bhaintwal, S. K. Wasan, *On quasi-cyclic codes over Z_q* AAECC, **20** 459-480, (2009).

[7] D. Boucher, W. Geiselmann, F. Ulmer, *Skew cyclic codes*, Appl. Algebra. Eng.Commun Comput., **18** 379-389, (2007).

[8] D. Boucher, P. Sole, F. Ulmer, *Skew constacyclic codes over Galois rings*, Advance of Mathematics of Communications, **2** 273-292, (2008).

[9] D. Boucher, F. Ulmer, *Coding with skew polynomial rings*, Journal of Symbolic Computation, **44** 1644-1656, (2009).

[10] A. R. Calderbank, E.M.Rains, P.M.Shor, N.J.A.Sloane, *Quantum error correction via codes over $GF(4)$* ,IEEE Trans. Inf. Theory,**44** 1369-1387, (1998).

[11] Y. Cengellenmis, A. Dertli, S.T. Dougherty, *Codes over an infinite family of rings with a Gray map*, Designs, Codes and Cryptography, **72** 559-580, (2014).

[12] A. Dertli, Y. Cengellenmis, S. Ere, *On quantum codes obtained from cyclic codes over A_2* , Int. J. Quantum Inform., **13** 1550031, (2015).

[13] A. Dertli, Y. Cengellenmis, S. Eren, *Quantum codes over the ring $F_2+uF_2+u^2F_2+...+u^mF_2$*, nt. Journal of Alg., **3** 115-121, (2015).

[14] J. Gao, *Skew cyclic codes over F_p+vF_p*, J. Appl. Math. & Informatics, **31** 337-342,(2013).

[15] J. Gao, L. Shen, F. W. Fu, *Skew generalized quasi-cyclic codes over finite fields*, arXiv: 1309.1621v1.

[16] M. Grassl, T. Beth, *On optimal quantum codes*, International Journal of Quantum Information, **2** 55-64,(2004).

[17] A. R. Hammons, V. Kumar, A. R. Calderbank, N. J. A. Sloane, P. Sole, *The Z_4-linearity of Kerdock, Preparata, Goethals and related codes*, IEEE Trans. Inf. Theory, **40** 301-319,(1994).

[18] S. Jitman, S. Ling, P. Udomkovanich, *Skew constacyclic codes over finite chain rings*, AIMS Journal.

[19] X.Kai, S.Zhu, *Quaternary construction bof quantum codes from cyclic codes over F_4+uF_4*, Int. J. Quantum Inform., **9** 689-700, (2011).

[20] S. Ling, P. Sole, *On the algebraic structures of quasi-cyclic codes I: finite fields*, IEEE Trans. Inf. Theory, **47** 2751-2760, (2001).

[21] S. Ling, P. Sole, *On the algebraic structures of quasi-cyclic codes II: chain rings*, Des.Codes Cryptogr., **30** 113130, (2003).

[22] S. Ling, P. Sole, *On the algebraic structures of quasi-cyclic codes III: generator theory*, IEEE Trans. Inf. Theory, **51** 2692-2000, (2005).

[23] Maheshanand, S. K. Wasan, *On Quasi-cyclic Codes over Integer Residue Rings*, AAECC, Lecture Notes in Computer Science, **4851** 330-336, (2007).

[24] J.Qian, *Quantum codes from cyclic codes over F_2+vF_2*, Journal of Inform.& computational Science **6** 1715-1722, (2013).

[25] J.Qian, W.Ma, W.Gou, *Quantum codes from cyclic codes over finite ring*, Int. J. Quantum Inform., **7** 1277-1283, (2009).

[26] J. F. Qian, L. N. Zhang, S. X. Zhu, *$(1+u)$-constacyclic and cyclic codes over F_2+uF_2*, Applied Mathematics Letters, **19** 820-823, (2006).

[27] I. Siap, T. Abualrub, N. Aydın, P. Seneviratne, *Skew cyclic codes of arbitrary length*, Int. Journal of Information and Coding Theory, (2010).

[28] P.W.Shor, *Scheme for reducing decoherence in quantum memory*, Phys. Rev. A., **52** 2493-2496, (1995).

[29] A. M. Steane, *Simple quantum error correcting codes* , Phys. Rev. A., **54** 4741-4751, (1996).

[30] M. Wu, *Skew cyclic and quasi-cyclic codes of arbitrary length over Galois rings*, International Journal of Algebra, **7** 803-807,(2013).

[31] X.Yin, W.Ma, *Gray Map And Quantum Codes Over The Ring $F_2+uF_2+u^2F_2$*, International Joint Conferences of IEEE TrustCom-11, (2011).

[32] S. Zhu, L. Wang, *A class of constacyclic codes over F_p+vF_p and their Gray images*, Discrete Math. **311** 2677-2682, (2011).

A New Hidden Web Crawling Approach

L.Saoudi
Computer science department
Mohammed Boudiaf University
M'sila, Algeria

A.Boukerram
Computer science department
Abderrahmane Mira University
bejaia, algeria

S.Mhamedi
Computer science department
Mohammed Boudiaf University
M'sila, Algeria

***Abstract*—Traditional search engines deal with the Surface Web which is a set of Web pages directly accessible through hyperlinks and ignores a large part of the Web called hidden Web which is a great amount of valuable information of online database which is "hidden" behind the query forms.**

To access to those information the crawler have to fill the forms with a valid data, for this reason we propose a new approach which use SQLI technique in order to find the most promising keywords of a specific domain for automatic form submission.

The effectiveness of proposed framework has been evaluated through experiments using real web sites and encouraging preliminary results were obtained

Keywords—Deep crawler; Hidden Web crawler; SQLI query; form submission; searchable forms

I. INTRODUCTION

The World Wide Web is a global information medium of interlinked hypertext documents accessed via computers connected to the internet. Most of the users rely on traditional search engines to search the information on the Web. These search engines deal with the Surface Web which is a set of Web pages directly accessible through hyperlinks and ignores a large part of the Web called hidden Web which is hidden to present-day search engines. It lies behind search forms and this part of the Web containing an almost endless amount of sources providing high quality information stored in specialized databases, only accessible through specific search interfaces created by using CGI and HTML forms or JavaScript etc [1]. which need to be filled manually by the user. A search interface consists of different form elements like text boxes, labels, buttons etc. User must provide an entry in at least one of them, and submit the form to obtain response pages containing the results of the query.

The hidden web crawler must also perform a similar filling process either by selecting suitable values from the domain of each finite form element or by automatically generating queries. The challenge is how to equip crawlers with the necessary input values for use in constructing search queries to obtain the optimized response pages without errors?

To address these challenges, we adopt a task-specific based SQLI approach to crawl the hidden Web.

The rest of the paper has been organized as follows: Section II describes different concepts related to hidden web crawler; section III describes the proposed work i.e. design of a Domain-specific hidden web crawler and explains the functionality of different components of crawler; section IV presents the progress of the experiment and its phases, section V describes the experimental results that is done over book domain which are discussed in section VI and finally, section VII draws the conclusion and describes the future research.

II. HIDDEN WEB CRAWLERS

A. Generality

A web crawler (also known as a robot or a spider) is a system for the bulk downloading of web pages. Web crawlers are used for a variety of purposes. Most prominently, they are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries. [2]

The whole web is divided into two types: the public web and the hidden web. The public web normally deploys by common use search engine, hidden web represents information, stored in specialized databases, only accessible through specific search interfaces created by using CGI and HTML forms or JavaScript etc. [3]

However, a number of recent studies [4,5,6,7] have observed that a significant fraction of Web content in fact lies outside the PIW(Publicly Indexable Web). Specifically, large portions of the Web are 'hidden' behind search forms, in searchable structured and unstructured databases (called the hidden Web or deep Web [6]). Pages in the hidden Web are dynamically generated in response to queries submitted via the search forms.

B. General Hidden web crawling strategy

The basic actions of a deep web crawler are similar to those of other traditional crawlers. A traditional web crawler selects URL's, retrieve pages, process the pages and extract links from the retrieved pages. The traditional crawlers do not distinguish between pages with and without forms[8]. Whereas, a Hidden web crawler performs additional sequence of actions for each form on a page [9]:

1) *Form detection:* the search form extractor looks for any <FORM> tags in the HTML web page to extract the associated search forms.

2) *Form Analysis:* Parse and process the form to build an internal form representation.

3) *Value assignment:* Use approximate string matching between the form labels and the labels in the database to generate a set of candidate value

4) *Form Submission:* send the filled form to the web server.

5) Response Analysis and Navigation: Analyze the response to check if the submission yielded valid search results. Crawling the hypertext links in the response page to some prespecified depth.

III. Our Proposed Approach

Our crawler is a domain specific Hidden Web crawler, Fig.1 depicts the proposed architecture of our crawler HiWC, having : a Web Page Analyzer, a Form structure and content classifier, the Form filler that uses a Domain Specific data repository and a response page analyzer.

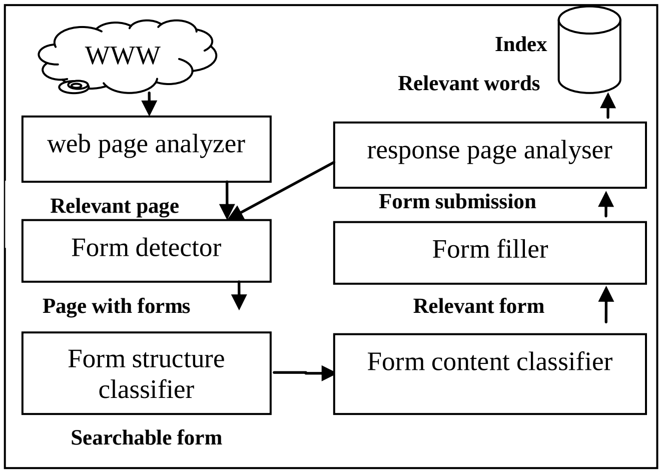


Fig. 1. general crawling process

Our crawling approach is divided in two phases:

A. *Domain Definition phase:*

In this phase (Fig2), we implement the domain definitions used to define a data-collection task. A domain definition is composed of two tables.

1) Labels table:
This table is an object relational table; it has three attributes label, alias and score. Each label *ai* has an associated name, a set of aliases {ai _alias1,…, ai _aliask}, and a specified score *si*.

A label represents a field that may appear in the search forms that are relevant to the domain.

The alias represents alternative labels that may identify the attribute in a query form. For instance, the attribute AUTHOR, from a domain used for collecting data about books, could have aliases such as “writer” or “written by”.

The score is a number between 0 and 1 which represent the weight of each label in the domain, for example the label ISBN should have the higher score in the book domain.

We collect search labels from training web sites, and then we find manually the alias of each label and give its score from its frequency in the search form in all the training sites.

2) Repository table:
This table has *m* attributes, each one take a value of the attribute label in the labels table (for example in the book domain, we take: ISBN, author, title...etc as attributes) we add an attribute which represent the weight of the object in its domain.

We obtain the initial repository values from 10 training search web sites.

we observed in several cases that the majority of search web page are vulnerable of SQLI queries, for this reason our crawler sends different SQLI queries to each search site to extract its response pages, for each obtained response page the crawler extract object values ,store its relevant information into database and give their weights , the weight is a value from 1 to 10, if the same object exist in the 10 web sites then its weight=10 which signify that this object is very important and should be among the first selected key words in the form submission process.

B. *Crawling phase:*

This phase passes through several steps as it shown in fig3:

1) Form detection
A standard HTML Web form consists of form tags, a start tag <form> and an end tag </form> within which the form fields reside. Form detector looks for any <FORM> tags in the HTML web page to extract the associated form.

2) Form structure classification
It was observed in several forms that there are content differences between searchable and non-searchable forms, For instance, forms with a password field, username, email or with a file upload field are non-searchable. The goal of the classification performed automatically was to exclude non-searchable forms.

3) Form content classification
To classify the search page into relevant or not relevant page, we extract form labels and use them in the classification process. The method we use to determine if a form is relevant to a domain consists of adding the frequency of each label, pondered by its predefined score, and checking if the sum exceeds the relevance threshold μ.

4) Form Submission
Once the system determines that a form is relevant to a certain domain d, the crawler select the promising objects with height weights to fill the form according to the matching process between the form labels and its equivalent on the labels table to provide a successful form submission.

5) Response page analysis
The response page to a form submission is received by a response analyzer module to distinguish between pages containing search results and pages containing error messages, different relevant values are extracted from successful response page. This feedback can be used to update the repository table. The obtained values are assigned scores that vary with time. The score of an object gets a positive (negative) boost ever time it is used in a successful (unsuccessful) form submission.

6) Indexing dynamic page
Indexes are data structures permitting rapid identification of which crawled pages contain like particular words or phrases, it has two types:

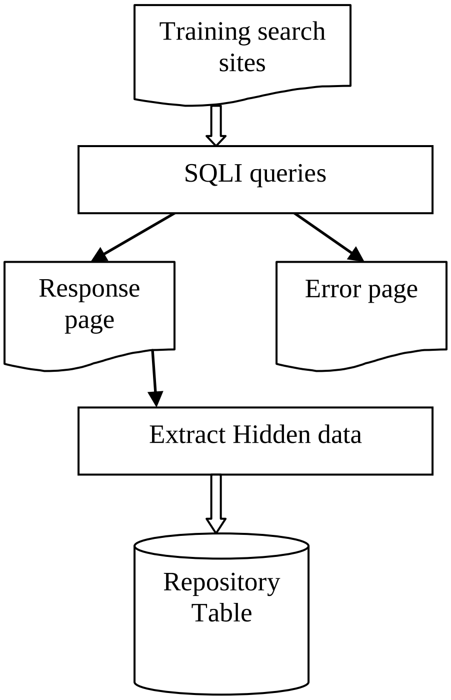


Fig. 2. Initial phase

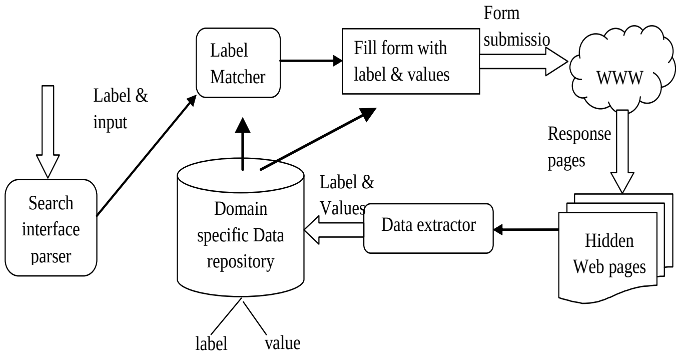


Fig. 3. crawling phase

- Inverted index: the inverted index stores a list of the URLs containing each word
- Forward index: The forward index stores a list of words for each URL

The rationale behind developing a forward index is that as URL are parsing, it is better to immediately store the words per URL. The forward index is sorted to transform it to an inverted index.

IV. Experimentation

To evaluate the performance of our approach, we test it on Books domain, our experimentation passes through two phases.

A. *Training phase:*

In this phase our crawler tries to collect initial data to implement the repository and labels tables.

The process for creating the domain definition was the following:

For book domain, we manually explored 10 sites at random as it shown in table1, and used them to define the attributes and its aliases. The specificity weight and the relevance threshold were also manually chosen from our experience visiting these sites.

TABLE I. Training Sites

Site name	**URL**	**SQLIA**
Book Depository	https://www.bookdepository.com/	query :)or'(1)=(1
ISBN Search	http://www.isbnsearch.org/	‘or’1’=’1
Paperback Swap	http://www.paperbackswap.com/	‘or’1’=’1
Tattered Cover Bookstore	http://www.tatteredcover.com/	‘or’1’=’1
AbeBooks	http://www.abebooks.com/	‘or’1’=’1
Free-eBooks.net	http://www.free-ebooks.net/	‘or’1’=’1
Green Apple Books&Music	http://www.greenapplebooks.com/	‘or’1’=’1
goodreads	https://www.goodreads.com/	‘or’1’=’1
Alibris	http://www.alibris.com/	‘or’1’=’1
Thomson Gale	http://www.cengage.com/	‘or’1’=’1

After having injected these sites, these latter will be analyzed for extracted all data of each book and put them in the table of labels that contains the label, aliases, and score as it shown in table2.

TABLE II. Labels Table

label	**alias**	**score**
Title	‘Name’, ‘title of book’	0.6
Author	‘By’, ‘Written by’, ‘Author’, ‘author’s name’	0.7
ISBN	ISBN-13, ISBN-10	0.95
Publisher	Editor	0.8
Format	‘binding type’	0.25
Category	Genre, sort	0.05
Price		0.05

B. *Testing phase:*

To evaluate the performance of our approach, we test it on Books domain. Once the domain was created, we use our crawler to crawl 12 websites called test websites. The list of websites visited by HiWC (Hideen Web Crawler) is shown in Table 3.

To check the accuracy of the obtained results, we manually analyzed the websites and compared the results with those obtained by our crawler. We measured the results at each stage of the process: associating texts with form fields, associating form fields with domain attributes, establishing the relevance of a form to a domain, and executing the queries on the relevant forms. To quantify the results, we used standard Information Retrieval metrics: precision, recall. The metrics defined to measure the performance of HiWC, make use of the following variables:

- FieldAttributeAHiWC: set of the associations between form fields and domain attributes discovered by HiWC.

- FieldAttributeAManual: set of the associations between form fields and domain attributes discovered by the manual analysis.

- FormDomainAHiWCt: set of the associations between forms and domains discovered by HiWC.

- FormDomainAManual: set of the associations between forms and domains discovered by manual analysis.

- SubmittedFormsHiWC: set of forms successfully submitted by HiWC.

TABLE III. TESTING SITES

Name	**URLs**
Blackwell's Bookshop	http://bookshop.blackwell.co.uk
The American Book Center	http://www.abc.nl
Strand Book Store	http://www.strandbooks.com
Dymocks Booksellers	https://www.dymocks.com.au
eCampus.com	http://www.ecampus.com
Powell's Books	http://www.powells.com
Barnes&Noble	http://www.barnesandnoble.com
Bookjetty	http://www.bookjetty.com
Listal	http://www.listal.com
Library thing	https://www.librarything.com/
BookFinder.com	http://www.bookfinder.com/
IWC Schaffhausen	http://www.iwc.com/en/collection/portugieser/

We defined the following metrics:

Metrics for associating labels and form fields.

PrecisionFieldAttributeA:= | FieldAttributeAHiWC ∩ FieldAttributeM | / | FieldAttributeAHiWC |

RecallFieldAttributeA = | FieldAttributeAHiWC ∩ FieldAttributeM | / | FieldAttributeM |

Metrics for Global associations between forms and domains:

Precision FormDomainA:= | FormDomainA HiWC ∩ FormDomainAM | / | FormDomainAHiWC |

RecallFormDomainA = | FormDomainAIHiWC ∩ FormDomainM | / | FormDomainM |

Precision SubmittedForms = |SubmittedForms HiWC | /| FormDomainA HiWC ∩ FormDomainAM|

V. EXPERIMENTAL RESULTS

In this section, we summarize some of the more significant results from these experiments.

We now take the 12 testing sites to crawl them and extract the forms, this latest be classified into two categories those which is searchable form and non-searchable form, as, it shown in table4.

TABLE IV. EXTRACTED FORMS CLASSIFICATION

Number of sites from which forms were picked	12
Total number of forms	25
Numbers of search forms for books	12

TABLE V. EXPERIMENTAL RESULTS

	D1(10)	D2(12)	D1+D2(22)
Submitted Forms			
Precision	13/13 1.00	12/12 1.00	25/25 1.00
Form-Domain Associations			
Precision	13/13 1.00	12/12 1.00	25/25 1.00
Recall	13/13 1.00	11/11 1.00	24/24 1.00
Field-Attribute Associations			
Precision	27/28 0.96	45/50 0.95	72/78 0.92
Recall	27/28 0.96	45/52 0.87	72/80 0.9

We performed a number of experiments to study and validate the overall architecture as well as the various techniques that we have employed.

Table V summarizes the obtained experimental results.

For each book domain, it shows the values obtained for all the metrics in the Training dataset (D1, the sites used to define the domains), the test dataset (D2, the testing sites) and in the Global dataset (D1+D2, Training+ testing).

It is important to notice that, in order to calculate the metrics for form-domain and field-attribute associations, "quick search" forms have not been considered.

VI. DISCUSSION

The obtained results are quite promising: all the metrics show high values and some of them even reach 1.00

Recall in associating forms and domains reached 1.00 in every case

The precision values obtained for the associations between attributes and form fields exceeded 0.92 with a recall 0.9.

The majority of the errors in this dataset came from a single source (Blackwell's Bookshop). If we did not have into account this source, the metrics would take values similar to those reached by the other ones.

VII. CONCLUSION

In this paper we described the conceptual and experimental study of the proposed approach. Our approach is based on a domain definition, which describe a data-collecting task based SQL injection technique to extract the most promising keywords of a specific domain for automatic form submission. We presented a simple operational model of a hidden Web crawler that succinctly describes the steps that a crawler must take: relevant page extraction, form detection, form structure classification, form content classification, form submission and response page analysis.

We described the architecture and design techniques used in HiWC, a prototype crawler implementation based on SQLI to get the initial keywords values and fill the repository database in the training phase. The promising experimental results using HiWC demonstrate the feasibility of hidden Web crawling and the effectiveness of our different techniques to implement this crawler.

In the future, we propose to handle forms powered by Javascript , that can significantly improve HiWC performance, and to test our crawler with different task specific domains.

REFERENCES

[1] L. Barbosa, J. Freire. “An Adaptive Crawler for Locating HiddenWeb Entry Points”, IW3C2 2007, Banff, Alberta, Canada, May 8–12, 2007.

[2] C. Olston, M. Najork, Web Crawling, now the essence of knowledge, vol. 4, No. 3 ,2010 ,pp. 175–246.

[3] K. K. Bhatia, A.K. Sharma, and R. Madaan, “AKSHR: A Novel Framework for a Domain-specific Hidden Web Crawler,” Proceedings of the 1st International Conference on Parallel, Distributed and Grid Computing , 2010, pp. 307–312.

[4] J. Madhavan, D. Ko L. Kot, “Google’s deep web crawl,” Proceedings of the 34th Internation Conference on Very Large Data Bases (VLDB), August, Auckland, New Zealand, (2008),pp. 1241-1252.

[5] M. Soulemane, M. Rafiuzzaman ,H. Mahmud, “Crawling the Hidden Web: An Approach to Dynamic Web Indexing” International Journal of Computer Applications.,vol. 55, No. 1,pp. 7-15, October 2012.

[6] G. Z. Kantorski, V. P. Moreira, and C. A. Heuser, “Automatic Filling of Hidden Web Forms: A Survey,” SIGMOD Record,Vol. 44, No. 1, , March 2015, pp. 44-35.

[7] B. Saharawat, A. Goyal, “Prefetching Data from Hidden Web with DSIM Architechture”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, No. 4, April 2013, pp. 887-890.

[8] S.Bal Gupta, “Challenges in designing a hidden web crawler”, International journal of information technilogy and system,vol.2,No.1, jan-jun 2013, pp.2277-9825

[9] S. Raghavan, H.Garcia-Molina, “Crawling the Hidden Web”, Technical Report 2000-36, Computer Science Department, Stanford University, December2000.

A Simulation Model for Nakagmi-m Fading Channel with m>1

Sandeep Sharma
School of ICT
Gautam Buddha University
Greater Noida, India

Rajesh Mishra
School of ICT
Gautam Buddha University
Greater Noida, India

Abstract—**In this paper, we propose a model to simulate a wireless fading channel based on Nakagami-m distribution with m>1. The Nakagami-m fading channel is the most generalized distribution as it can generate one-sided Gaussian distribution, Rayleigh distribution and Rician distribution for m equals to 0.5, 1 and >1 respectively. In this work we have proposed a method to generate a wireless fading channel based on Nakagami-m distribution as this distribution fits to a wide class of fading channel conditions. Simulation results were obtained using Matlab R2013a and compared with the analytical results.**

Keywords—Nakagami Distribution; Fading Channel; Wireless Channel Modeling

I. INTRODUCTION

In wireless communication fading plays a vital role in the channel estimation. Fading is the rapid fluctuation in the received signal strength of the wireless signal. Communication systems are subjected to fading caused by multipath propagation due to reflections by surrounding objects, refractions and scattering by buildings and other large structures. Thus, the received signal is a sum of various signals that arrive at the receiver via different propagation paths which may be direct line of sight (LOS) or non line of sight path (non-LOS). To model fading in wireless communication, several techniques have been used in literature. As the nature of the wireless channel is random, it has to be model statistically. Several statistical models have been used in the literature to describe the fading envelope of the received signal [6],[13]-[15].

The Rayleigh and Rician distributions are used to characterize the fading envelope of the wireless signals over small geographical areas or short term fades while the log-normal distribution is used when much wider geographical areas are involved. A more versatile statistical model, however, is Nakagami’s m-distribution [1], which can model a variety of fading environments including those modeled by the Rayleigh and one-sided Gaussian distributions. Also the log-normal and Rician distributions may be closely approximated by the Nakagami distribution in some ranges of mean signal values [l6]. The fit between Nakagami and Rician distributions is very accurate for low signal-to-noise ratio (SNR) values in comparison to large SNR values. Furthermore, the Nakagami distribution is more flexible and more accurately fit experimental data for many physical propagation channels then the log-normal and Rician distributions [l6],[17]. We may find various research papers where Nakagami distribution is used to simulate in applications like satellite communication, vehicular to vehicular communication, even it is applied in medical applications such as ECG and ultrasound signals. Although the Nakagami model fits experimental data around the mean or median, but it is reported in [18] that it does not fit very well in the tails of the distribution. In spite this, the Nakagami distribution is much popular and used by many of the researchers in their domain whether it may be wireless, medical, terrestrial signal analysis, and vehicular ad-hoc networks.

Paper Organization: This paper is organized as follows. In section II, we discussed the theoretical background of wireless channel modeling and the factors affecting it. Section III explains the Nakagami-m distribution followed by the simulation method in section IV. Section V discusses the various results and their analysis. Finally, section VI.

II. THEORITICAL BACKGROUND

In this section, we explain theoretical background of a wireless channel and the factors affecting the channel response. A wireless channel is different from a wired channel as it contains multipath components from direct line of sight (LOS) component and various reflected and refracted components. The wireless channel is made by the constructive and destructive addition of different multipath components introduced by the channel. The same phase components are added and the out of phase components are subtracted and their algebraic sum is what we get at the antenna of the receiver. In general, the deterministic channel models are rarely available as the nature of the channel is random, and thus we need to characterize multipath channels statistically. If a single pulse is transmitted over a multipath channel, then the received signal will not be a single pulse but appear as a series of pulses, with each pulse in the series corresponds to the LOS component or an individual multipath component associated with a discrete scatterer or cluster of scatterers. The channel characteristics certainly depends upon the number of scatterer objects, number of multipath, size of the objects and the amount of absorption by the surrounding environment such as wall and roofs (thickness and material has a impact on the degree of absorption). Another characteristic of the multipath channel is its time-varying nature. This variation in time arises because either the transmitter or the receiver is moving, and this mobility of the transmitter and/ or receiver therefore change the location of reflectors in the transmission path, which give rise to multipath, will change over time.

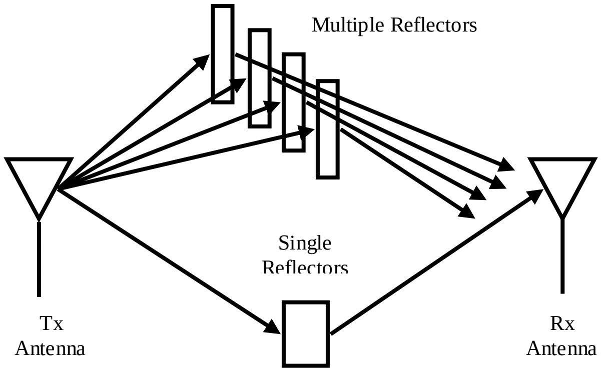


Fig. 1. Multipath due to single and multiple reflectors

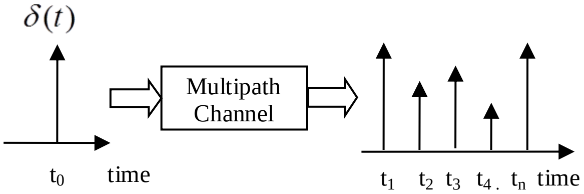


Fig. 2. Effect of Multipath when only a single impulse is transmitted

Another important characteristic of a multipath channel is the time delay spread which is caused to the received signal. This delay spread equals the time of arrival of the first received signal component (LOS or multipath) and the last received signal component associated with a single transmitted pulse. When the delay spread is small compared to the inverse of the signal transmission bandwidth then there is little time spreading in the received signal. However, when the delay spread is relatively large, there is considerable time spreading of the received signal which can lead to signal distortion substantially.

A. Envelope and Power Distribution in the Signal

Consider any two Gaussian random variables X and Y, both with zero mean and equal variance σ2, then resultant $R=\sqrt{X^2+Y^2}$ of X and Y can be related by Rayleigh-distributed and Z2 is exponentially distributed. If r_I and r_Q represents the in-phase and quadrature phase components having the variance of σ2, then the envelope of the signal is given by

$$z(t)=|r(t)|=\sqrt{r_I^2+r_Q^2} \tag{1}$$

and it is Rayleigh distributed by the well known distribution given by

$$p_Z(z)=\frac{2z}{P_r}\exp\left[-\frac{z^2}{P_r}\right]=\frac{z}{\sigma^2}\exp\left[-\frac{z^2}{2\sigma^2}\right],\quad x\geq 0, \tag{2}$$

where, $P_r=\sum_n E\left[\alpha_n^2\right]=2\sigma^2$ is the average received power based on the path loss and shadowing alone. Making substitution $z^2(t)=|r(t)|^2$ in the Rayleigh distribution we get:

$$p_{Z^2}(x)=\frac{1}{P_r}e^{-\left(\frac{x}{P_r}\right)}=\frac{1}{2\sigma^2}e^{-\left(\frac{x}{2\sigma^2}\right)},\qquad x\geq 0 \tag{3}$$

This shows that the received power is exponentially distributed over mean $2\sigma^2$ .So $r(t)$ has a Rayleigh-distributed amplitude and uniform phase, and the two are mutually independent. When the channel has a fixed line of sight (LOS) component, then $r_I(t)$ and $r_Q(t)$ do not have zero-mean. In such a case, the received signal is the superposition of a complex Gaussian component and a LOS component. The signal envelope in this case can be shown to have a Rician distribution and given by:

$$p_Z(z)=\frac{z}{\sigma^2}\exp\left[\frac{-(z^2+s^2)}{2\sigma^2}\right]I_0(\frac{zs}{\sigma^2}),\quad x\geq 0, \tag{4}$$

where, $2\sigma^2=\sum_{n,n\neq 0}E\left[\alpha_n^2\right]$ is the average amount of power in the non-line of sight component and $s^2=\alpha_0^2$ is the average amount of power in the line of sight (LOS) component of the radio signal. Here I_0 is the modified Bessel's function of zero order. The Rician fading has the average received power which is given by the equation:

$$P_r=\int_0^{\infty} z^2 pz(z)dx=s^2+2\sigma^2 \tag{5}$$

Very often, the Rice distribution is described in terms of fading parameter "K" defined as

$$K=\frac{s^2}{2\sigma^2}=\frac{Power\,in\,the\,LOS\,component}{Power\,in\,the\,non-LOS\,multipath\,component} \tag{6}$$

Different values of K give us different fading statistics and thus it is a factor that controls the amount of fading in the wireless channel. For K = 0 we have Rayleigh fading, and for K = 1 we have no fading, i.e. a channel with no multipath and we have a LOS component. The fading parameter "K" is therefore a measure of the severity of the fading: a small "K" implies severe fading, a large value of K implies low fading.

$$K=\begin{cases}0 & Rayleign\,Fading\\ 1 & No\,Fading\,only\,LOS\end{cases} \tag{7}$$

If we substitute $s^2=KP/K+1$ and $2\sigma^2=P/K+1$ the Rician distribution can be obtain in terms of K

$$p_Z(z)=\frac{2z(K+1)}{P_r}\exp\left[-K-\frac{(K+1)z^2}{P_r}\right]I_0\left(2z\sqrt{\frac{K(K+1)}{P_r}}\right),z\geq 0 \tag{8}$$

The more general distribution is the Nakagami distribution with the help of which we can generate both the Rayleigh fading as well as Rician fading.

III. The Nakagami Distribution

With Nakagami-m distribution [1], usually denoted by m-distribution, a wide range of fading channel conditions can be modeled.

This fading distribution has often provides the best fit to land-mobile, indoor mobile multipath propagation as well as for the ionospheric radio links [2]. Recent studies also showed that Nakagami-m gives the best fit for satellite-to-indoor and satellite-to-outdoor radio wave propagation as well [3, 4]. The channel response of a wireless channel is a complex quantity and for this reason let us assumes that the complex valued Nakagami-m fading channel Z is represented as follows:

$$Z = X + jY = \mathrm{R}\left(e^{j\Theta}\right) \tag{9}$$

where X, Y, R and Θ represents the in-phase component, the quadrature-phase component, the envelope and the phase component simultaneously. The probability density function (PDF) for a Nakagami-m distributed fading envelope R can be expressed as:

$$f_R(r) = \frac{2m^m r^{2m-1}}{\Gamma(m)\Omega^m}\exp(-\frac{mr^2}{\Omega}), \quad 0 \le r < \infty \tag{10}$$

Where, Ω=E[R2], is the expected value of the average power and m is shaping parameter which controls the shape of the distribution. When m is integer, R is the square root of the sum of the amplitude square of m i.i.d. complex GaussianRVs,

$$R = \sqrt{\left|x_1^{\,2}\right| + \left|x_2^{\,2}\right| + \left|x_3^{\,2}\right| + \cdots \left|x_m^{\,2}\right|} \tag{11}$$

Where x_i = 1,2,3…….m is a complex Gaussian random variable (RV) with zero mean and variance Ω/m. Here E(.) is the expectation operator and Γ(.) is the gamma function. m is the inverse of the normalized variance of R2:

$$m = \frac{\left(E\left[R^2\right]\right)^2}{Var(R^2)} = \frac{\Omega^2}{Var(R^2)} \tag{12}$$

where $Var(R^2)$ is the variance of R^2 .The value for m ranges between 1/2 and ∞ .When $m \to \infty$, the channel converges a static channel i.e. it no longer remains variant channel.[5]. As special cases, for $m = 1$ the Nakagami-m become Rayleigh distribution one-sided Gaussian distribution for $m = 1/2$. This principally means that, fading is more severe than Rayleigh fading if m < 1, and for values of m > 1, the fading is less severe. For the values of m > 1, the Nakagami-m distribution closely approximates the Rician distribution. The Nakagami shape parameter m and the Rician factor K which determines the severity of fading in case of the Rician fading can be related to the following equation [5]:

$$m = \frac{\left(K+1\right)^2}{2K+1} \text{ ,for K}\ge 0 \tag{13}$$

$$m = \begin{cases} m = 1/2; \textit{One Sided Gaussian Distribution} \\ m = 1; \qquad \textit{Rayleigh Distribution} \\ m > 1; \quad \textit{Rician Distribution} \quad , m = \dfrac{(K+1)^2}{(2K+1)} \\ m = \infty; \qquad \textit{No Fading} \end{cases}$$

The Nakagami-m phase envelope joint distribution is given by

$$f_{R\Theta}(r,\theta) = \frac{m^m \left|\sin(2\theta)\right|^{m-1} r^{2m-1}}{2^{m-1}\Omega^m \Gamma^2\left(\dfrac{m}{2}\right)}\exp(-\frac{mr^2}{\Omega}) \tag{14}$$

The envelope pdf $f_R(r)$ is given by the well known formula

$$f_R(r) = \frac{2m^m r^{2m-1}}{\Gamma(m)\Omega^m}\exp(-\frac{mr^2}{\Omega}), 0 \le r < \infty \tag{15}$$

And the phase envelope is given by:

$$f_\Theta\left(\theta\right) = \frac{\Gamma\left(m\right)\left|Sin2\theta\right|^{m-1}}{2^m \Gamma^2\left(m/2\right)}, \quad -\pi \le \theta < \pi \tag{16}$$

$$f_X\left(u\right) = f_Y\left(u\right) = \frac{m^{(m/2)}\left|u\right|^{m-1}}{\Omega^{(m/2)}\Gamma\left(m/2\right)}\exp\left(-\frac{mu^2}{\Omega}\right), -\infty < u < \infty \tag{17}$$

IV. Simulation Method

A pair of correlated Nakagami fading envelopes can be generated keeping the integer value of the shaping or fading parameter, from m-dimensional i.i.d. (independent and identically distributed) complex Gaussian distributed column vector $\bar{Z}_1$ and $\bar{Z}_2$ i.e.

$$\bar{Z}_1 = \left[x_1 \;\; x_2 \;\; x_3 \;\; \ldots x_m \;\; \right]^T \text{and} \bar{V}_1 = \left|\bar{Z}_1\right|^2 \tag{18}$$

$$\bar{Z}_2 = \left[y_1 \;\; y_2 \;\; y_3 \;\; \ldots y_m \;\; \right]^T \text{ and } \bar{V}_1 = \left|\bar{Z}_1\right|^2 \tag{19}$$

$$x_i = x_{ireal} + x_{iim} \quad \text{and} \quad y_i = y_{ireal} + y_{iim} \tag{20}$$

where $x_1 , x_2 \cdots x_m$ and $y_1 , y_2 \cdots y_m$ are complex Gaussian RVs with zero mean and variance $\sigma_x^{\,2}$ and $\sigma_y^{\,2}$ respectively. We assume that the power in the real part and imaginary part be equal in magnitude. Here, $x_{ireal} , y_{ireal} , x_{iim} , y_{iim}$ have normalized envelope correlation coefficients ρ_{xy} and its range lies between 0 and 1, i.e. ($0 \le \left|\rho_{xy}\right| \le 1$) and at the same time the real and imaginary parts of the random variables x_i and y_i remains uncorrelated. In eqn 18 and eqn 19, transpose is denoted by T. Here, ρ is the correlation of power between $\bar{V}_1 \; and \; \bar{V}_2$ and related as follows:

$$\rho = \left|\rho_{xy}\right|^2 = \frac{\mathrm{cov}\left(\bar{V}_1, \bar{V}_2\right)}{\sqrt{\mathrm{var}\left(\bar{V}_1\right).\mathrm{var}\left(\bar{V}_2\right)}} \tag{21}$$

Where, $\mathrm{var}\left(\bar{V}_1\right) = \Omega, \; \mathrm{var}\left(\bar{V}_1\right) = \hat{\Omega}$ and$\left(0 \le \rho \le 1\right)$. The joint pdf of $\sqrt{\bar{V}_1} \, and \, \sqrt{\bar{V}_2}$ gives the joint pdf of the nakagami-m distribution.

The following steps are involved in the generation of the Nakagmi-m fading channel

Step1: First generate m i.i.d. Gaussian random variables.

Step2: Generate vectors $\bar{Z}_1$ and $\bar{Z}_2$ using eqn 18 and eqn 19.

Step3: Calculate var , cov and ρ using eqn 21.

Step4: Generate complex channel coefficients as fig. 3.

Step5: Initialize k and calculate m using eqn 13.

Step6: Generate the channel using eqn 14, eqn 15 and eqn 16.

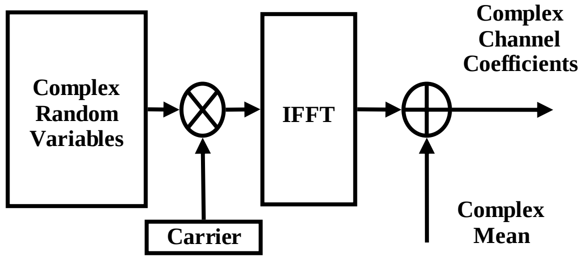


Fig. 3. Generation of the Channel Coefficients

V. Results Analysis

We have simulated Rayleigh, Rician and Nakagami-m fading channel. There is a relationship between the fading factors ‘K’ and ‘m’ as shown in the eqn 13, so we have initialize ‘K’ and evaluated ‘m’ the fading factor of Nakagami-m fading channel. The simulation is done using Matlab R2013b and the results are shown for various value of the fading parameter m. Figure 4 shows the Nakagami-m distribution for same value of scaling parameter Ω and various value of the fading parameter m. Figure 5 shows the effect of the scaling parameter Ω, keeping the fading parameter m as constant. In the fig. 6 and 7 the simulated Rayleigh and Rician fading channel were simulated for m=1. As per the eqn. 10, for m=1, the Nakagami-m distribution converts into a Rayleigh distribution, and the Rician distribution is merely an impulse which could be verified by fig.7 as shown. In the fig.8, the simulated channel is compared with the theoretical channel for m=1. In the subsequent figures, we have shown the Rayleigh, Rician and Nakagami-m channel for various values of m as shown in the figure. It is interesting to know that when the value of m>1, the simulated Nakagami-m channel follows Rayleigh fading distribution which could be verified by our simulation results. The channel coefficients are complex values having a real part and imaginary part and it is plotted for absolute amplitude against 1000 samples and is shown in the fig.21. In this work, we have also find out the impulse response of the channel by giving the channel input as 000000000010000000000 and 00000100000 as shown by fig. 23 and fig.22 respectively.

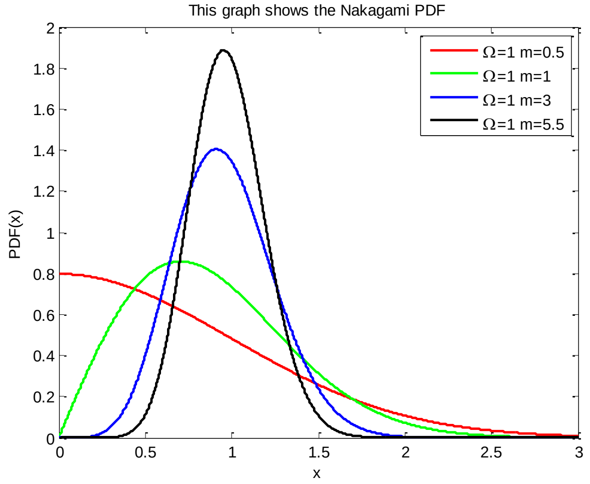


Fig. 4. Nakagami-m distribution plot for Ω=1 and m=0.5,1,3,5.5

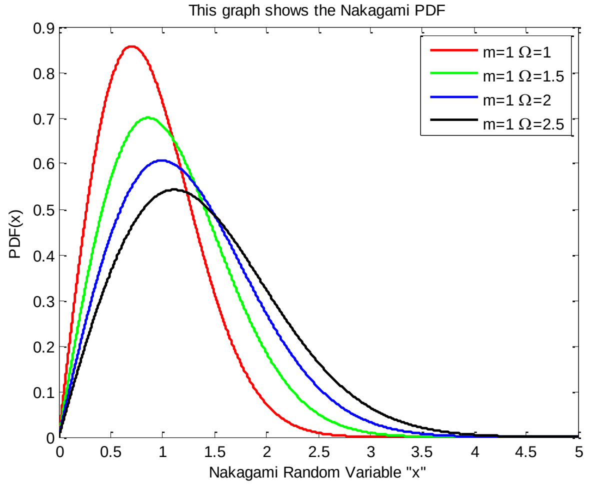


Fig. 5. Nakagami-m distribution plot for m=1 and Ω=1,1.5,2,2.5

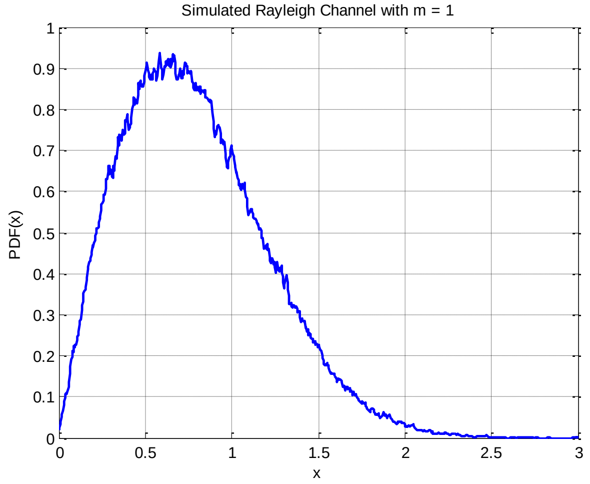


Fig. 6. Simulated Rayleigh Fading Channel with m=1

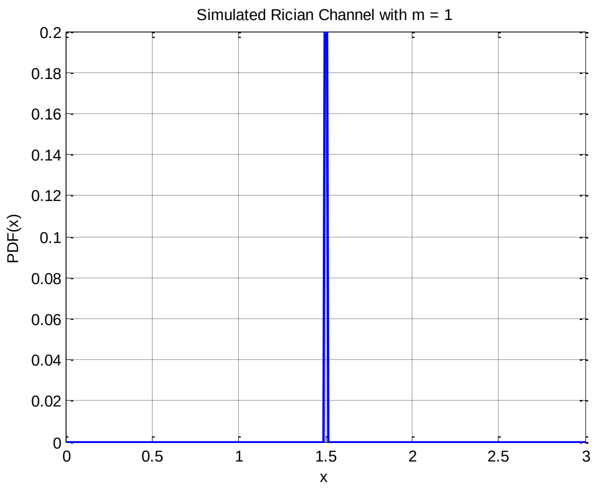


Fig. 7. Simulated Rician Fading Channel with m=1

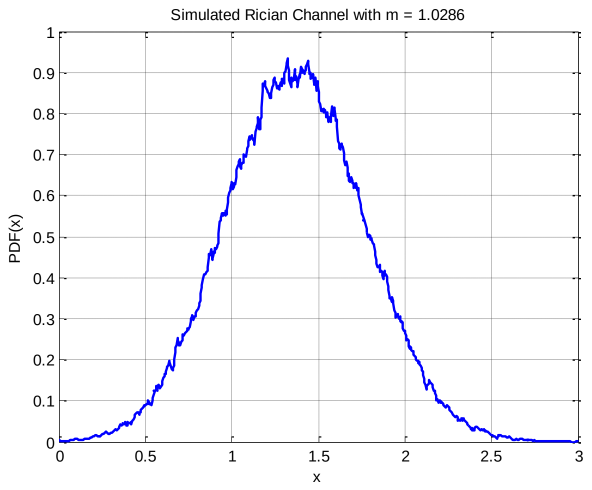


Fig. 10. Simulated Rician Fading Channel with m=1.0286

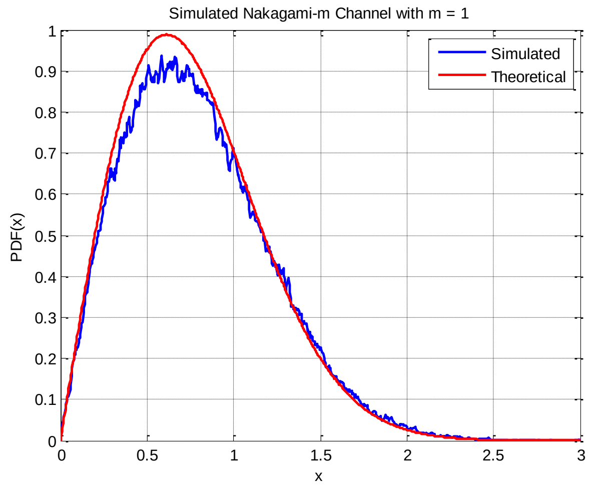


Fig. 8. Simulated Nakagami-m Fading Channel with m=1

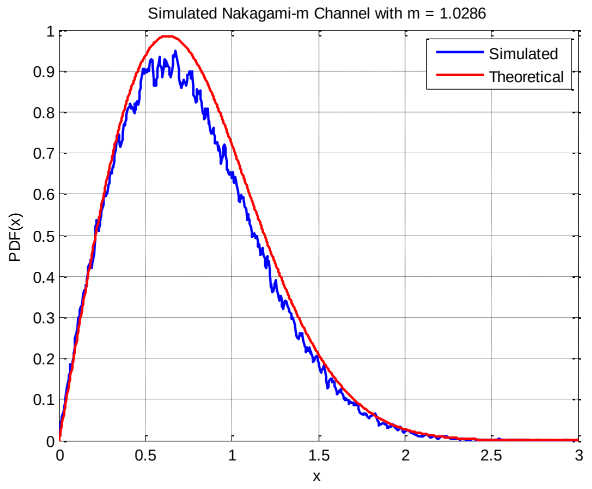


Fig. 11. Simulated Nakagami-m Fading Channel with m=1.0286

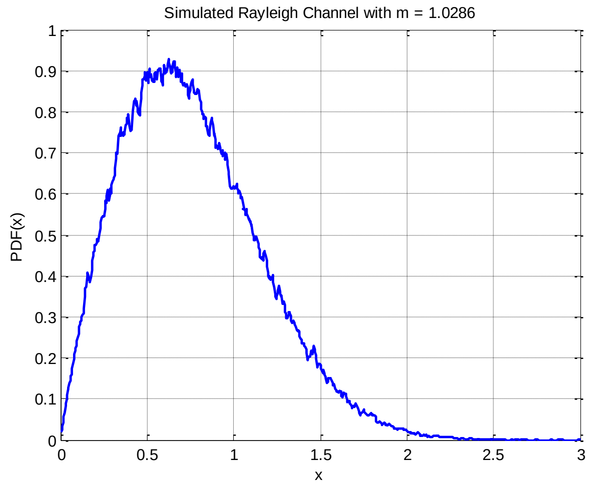


Fig. 9. Simulated Rayleigh Fading Channel with m=1.0286

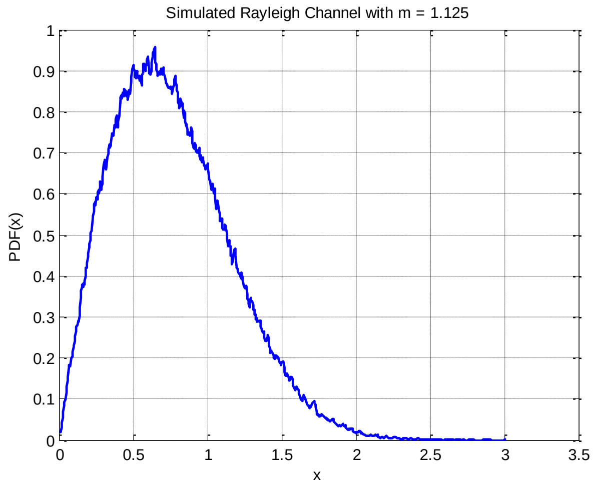


Fig. 12. Simulated Rayleigh Fading Channel with m=1.125

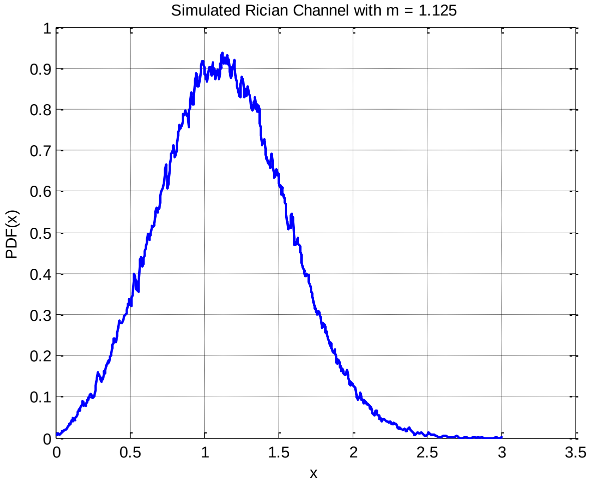


Fig. 13. Simulated Rician Fading Channel with m=1.125

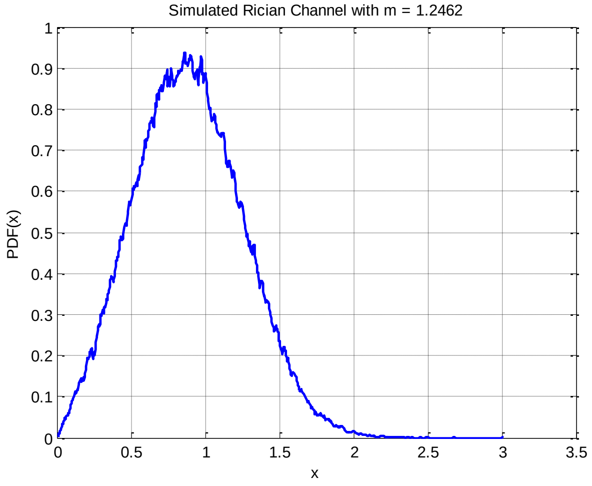


Fig. 16. Simulated Rician Fading Channel with m=1.2462

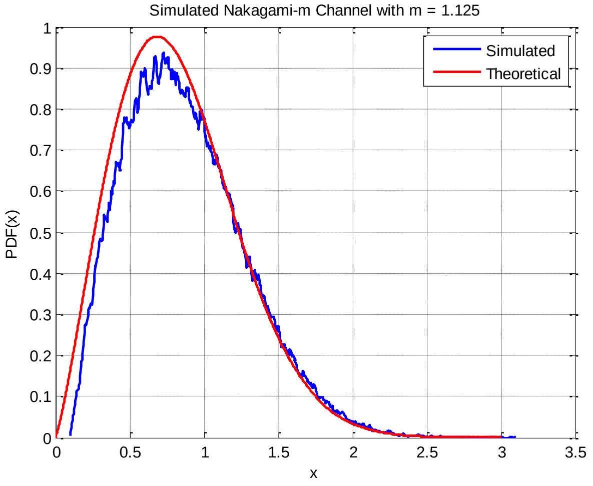


Fig. 14. Simulated Nakagami-m Fading Channel with m=1.125

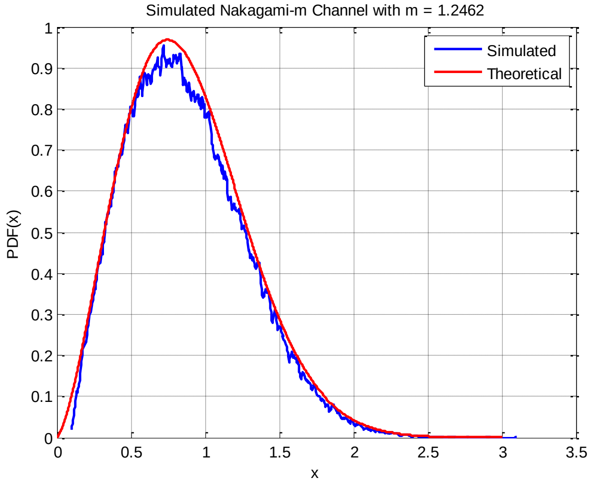


Fig. 17. Simulated Nakagami-m Fading Channel with m=1.2462

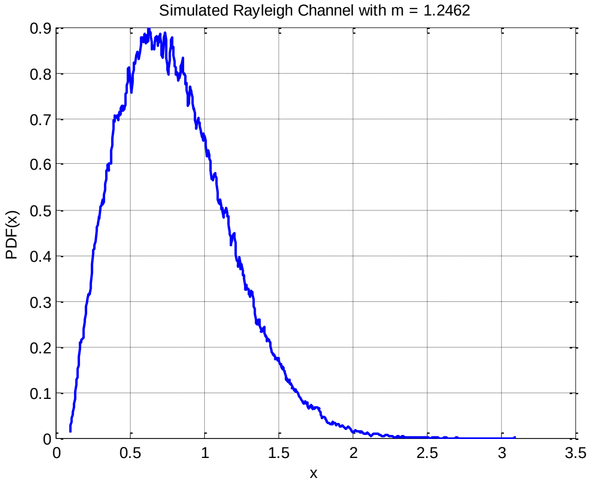


Fig. 15. Simulated Rayleigh Fading Channel with m=1.2462

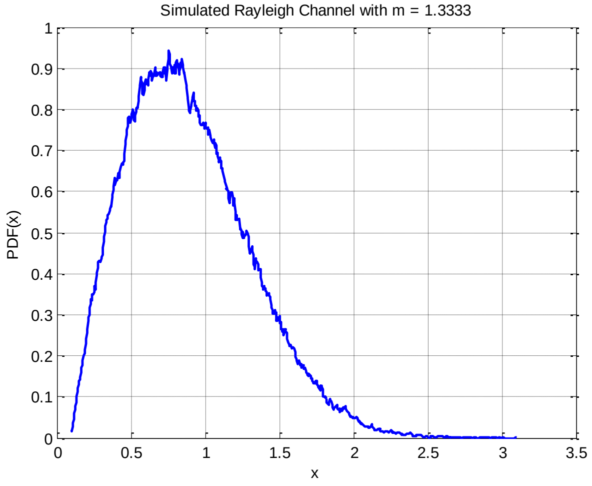


Fig. 18. Simulated Rayleigh Fading Channel with m=1.3333

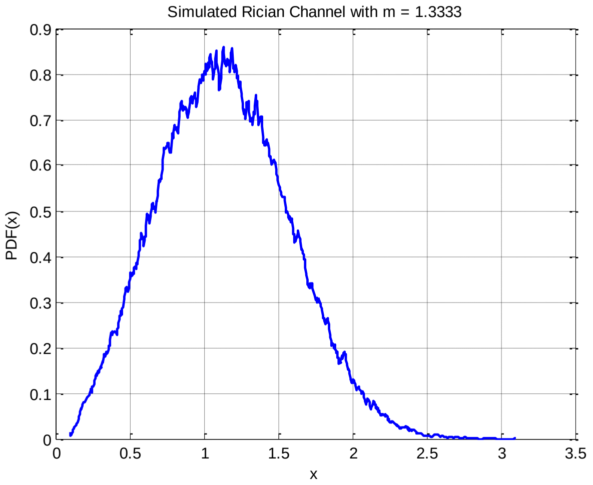


Fig. 19. Simulated Rician Fading Channel with m=1.3333

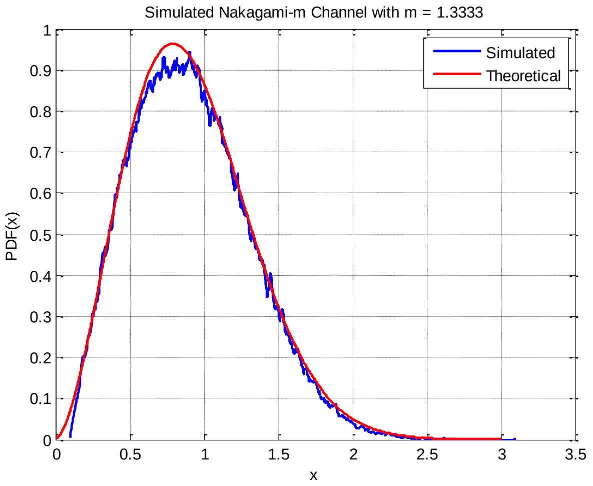


Fig. 20. Simulated Nakagami-m Fading Channel with m=1.3333

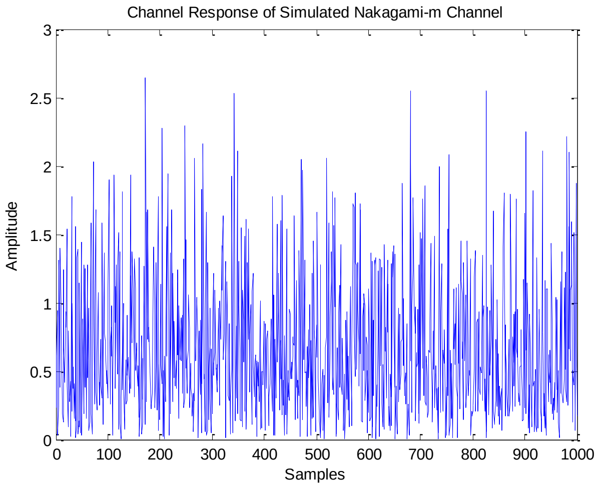


Fig. 21. Channel Response of the Simulated Nakagami-m Fading Channel

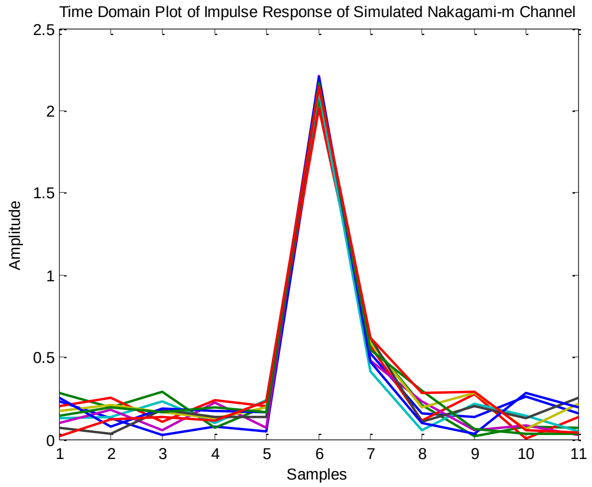


Fig. 22. Impulse Response of the Simulated Nakagami-m Fading Channel when channel input is 00000100000

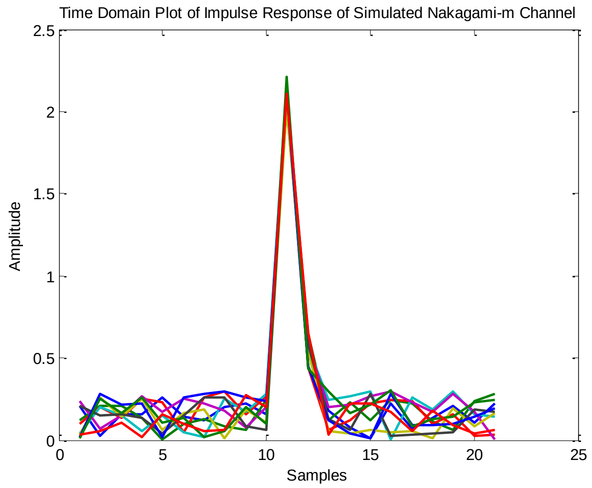


Fig. 23. Impulse Response of the Simulated Nakagami-m Fading Channel when channel input is 000000000010000000000

VI. Conclusions and Future Work

In this paper, we have discussed the theory needed for the simulation of Nakagami-m fading channel, and propose a statistical model to simulate the same. Rayleigh, Rician and Nakagmai-m fading channels have been simulated and the result of Nakagmi-m channel simulated with the theoretical. In this work, channel response and channel impulse response were also calculated. In future, Monte-Carlo simulation could be applied to know the standard deviation of channel coefficients thus generated over a number of iteration.

References

[1] M. Nakagami. The *m*-distribution - A General Formula of Intensity Distribution of Rapid Fading.In *W. C. Hoffman: Statistical Methods of Radio Wave Propagation*, Oxford, England, 1960.

[2] M.K. Simon, J.K. Omura, R.A. Scholtz, and B.K. Levitt. Spread Spectrum Communication Handbook. McGraw-Hill Inc, New York, revised edition, 2002.

[3] A. Lakhzouri, E. S. Lohan, I. Saastamoinen, and M. Renfors. “Measurement and Characterization of Satellite-to-Indoor RadioWave Propagation Channel”. In CDROM Proc. of The European Navigation Conference (ENC-GNSS ’05), Munich, Germany, Jul 2005.

[4] A. Lakhzouri, E. S. Lohan, I. Saastamoinen, and M. Renfors. “Interference and Indoor ChannelPropagation Modeling Based on GPS Satellite Signal Measurements”. In Proc. of ION GPS, pages 896–901, Sep 2005.

[5] M. D. Yacoub, G. Fraidenraich, and J. C. S. Santos Filho, “Nakagami-m phase-envelope joint distribution,” Electron. Lett., vol. 41, no. 5, pp.259–261, Mar. 2005.

[6] A.F. Abouraddy and S.M. Elnoubi. “Statistical Modelling of the Indoor Radio Channel at 10 GHz through Propagation Measurements - Part 1: Narrow-band Measurements and Modelling”. IEEE Trans. on Vehicular Technology, 49(5):1491–1507, Sep 2000.

[7] M.K. Simon and M.S. Alouini. Digital Communications over Fading Channels: A Unified Approach to Performance Analysis. Wiley InterScience, Sep 2000.

[8] N. C. Beaulieu and C. Cheng. “An Efficient Procedure for Nakagami-m Fading Simulation”. In IEEE Proc. of Globecom 2001, volume 6, pages 3336–3342, Nov 2001.

[9] K. W. Yip and T. S. Ng.,”A simulation model for nakagami-m fading channels, m < 1”. IEEETransactions on Communications, 48(2):pp. 214–221, Feb 2000.

[10] L. Schumacher, J. P. Kermoal, F. Frederiksen, K. I. Pedersen, A. Algans, and P. Mogensen. MIMO Channel Characterisation. IST Project IST-1999-11729 METRA Deliverable D2, Feb 2001.

[11] K.I. Pedersen, J. B. Andersen, J. P. Kermoal, and P. Mogensen. “A stochastic multiple-input-multipleoutput radio channel model for evaluation of space-time coding algorithms”. In Proc. of IEEE VTC Fall, volume 2, pages 893–897, Boston, USA, Sep 2000.

[12] L. Schumacher, K. I. Pedersen, and P. E. Mogensen. “From antenna spacings to theoretical capacities-guidelines for simulating MIMO systems” . In 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, volume 2, pages 587–592, Sep 2002.

[13] K.N. Nagesh, D. Satyanarayana, S.M.Prabhu, “ Statistical analysis of MIMO scheme under Nakagami fading channels”,in the proc. of IEEE 16th International Conference on Advanced Communication Technology(ICACT),pp.255 -259, 2014.

[14] Hussain, S.; Fernando, X.N, “Performance Analysis of Relay-Based Cooperative Spectrum Sensing in Cognitive Radio Networks Over Non-Identical Nakagami- m Channels”, IEEE Transactions on Communications, vol.62, issue 8, pp.2733 - 2746, 2014.

[15] JianxiaLuo; Zeidler,J.R.“A statistical simulation model for correlated Nakagami fading channels”, in the proc. of IEEE International Conference on Communication Technology Proceedings, 2000. WCC - ICCT 2000, vol.2, pp.1680 – 1684.

[16] H. Suzuki, “A statistical model for urban radio propagation,” IEEE Trans.on Commun., vol. COM-25, pp. 673-680, July 1977.

[17] U. Charash, “Reception through Nakagami fading multipath channels with random delays,” fEEE Trans. Commun., vol. COM-27, pp. 657470, Apr. 1979.

[18] S . Stein, “Fading channel issues in system engineering,” IEEEJ. Selecr. Areas Commun., vol. SAC-5, pp. 6849, Feb. 1987.