International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 115 - 125, 2016, ISSN:- 2319–8656

Security Requirements and Security Threats In Layers Cloud and Security Issues Open Source Cloud

Roya Morshedi
Department of security
Information Engineering
Central branch
University of Malek ashtar
Tehran, Iran
Royamorshedi@gmail.com

Ali Payandeh
Department of Infrmation and
Communication Tecnology,ICT
Central branch
University of Malek ashtar
Tehran, Iran

Ali Pourghaffari
Department of Infrmation and
Communication Tecnology,ICT
Central branch
University of Malek ashtar
Tehran, Iran

Abstract: Euacalyptus, OpenNebula and Nimbus are three major open-source cloud-computing software platforms. The overall function of these systems is to manage the provisioning of virtual machines for a cloud providing infrastructure-as-a-service. These various open-source projects provide an important alternative for those who do not wish to use a commercially provide cloud. This is a fundamental concept in cloud computing, providing resources to deliver infrastructure as a service cloud customers, making users have to buy and maintain computing resources and storage. In other hand, cloud service providers to provide better resources and facilities customers need to know they are using cloud infrastructure services. In this end, we intend to security threats in the cloud layer and then to analyse the security services in cloud computing infrastructure as a service to pay.

Keywords: Infrastructure as a Service, Infrastructure as a Service security threats, security issues cloud computing infrastructure services

1. INTRODUCTION

Cloud computing infrastructure has unique properties compared to other layers of the source specification introduced new security risks to the community and security experts, industry practitioners and experts to find solutions appropriate security characteristics and risks of this new trend.Cloud computing is a relatively recent concept which combines technologies for resource management and provisioning with the ideas of mass deployment, elasticity and ease of use. To enterprises it is an interesting concept on several levels – from internal applications to the possibility of sharing resources with other organization or providing their own resources as a service to others. Predictions by IDC Adriatics suggests that 2011 will be a year of transition for the global cloud computing services as it is expected that the related technologies will graduate from the early adoption to the new mainstream phase [5]. Cloud computing has found significant support in the business world, with expected rises in the revenue coming from cloud-related services as high as 30%, public clouds valued at USD 29 billion, and private clouds valued at USD 13 billion. Predictions for more distant future are even more optimistic, with some predictions of its growth by 2014 being notably higher (as much as up to five times) than the average global IT spending, with a compound annual growth rate of 27%. Enterprises have a number of reasons to adopt cloud computing technologies among which are [4][10]: easier management of their resources, introduction of dynamic infrastructure, per-consumption billing, support for varied platforms and operating systems and the possibility to start and stop the provisioned resources as needed.

With the spread of computers, scientists are exploring ways to solve that increased computing power, processing power, resilience and optimal use of infrastructure, platforms and applications were discussed. The first scientific use of the term cloud computing was in an article in 1997. Amazon modernization of data centers, cloud computing has played a key role in the development and Eucalyptus in early 2008, the first open source platform for deploying private clouds became AWS- API compatible. In early 2008, OpenNebula, open source software for deploying private and hybrid clouds and for the federation of clouds. Despite all the benefits of this environment, there are security concerns in two main groups, security, cloud service providers and cloud customers are security issues. In this article we are going to discuss security issues in cloud infrastructure services.

2. CLOUD COMPUTING DEFINED

The basic concept of cloud computing and initiator Name of the 1950s, when large-scale mainframe computers in universities and companies through terminal was available. For efficient use of such processors, it is recommended that users can access these computers simultaneously from multiple terminals

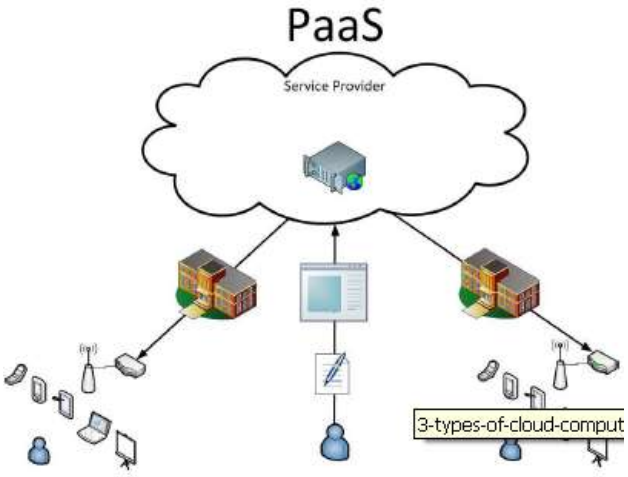
to share your information. With extensive computer scientists to explore solutions that enhance the computing power, processing power, high resilience and optimum use of infrastructure, platforms and applications were discussed. The first practical use of the term cloud computing in an article in 1997. Amazon modernization of data centers, cloud computing has played a key role in the development and Eucalyptus in early 2008, the first AWS- API compatible platform for deploying private clouds became open source. In early 2008, OpenNebula, open source software for deploying private and hybrid clouds as well as for the Federation of clouds.

They directly lead to introduction of the three common service models in which cloud computing is implemented [3]:

• *Software as a Service (SaaS)*, where the product is an application (usually a Web application) offered to users with little to no customizations. The users may have high-level administrative access to the application but have no control or influence on the application's implementation, inner workings or underlying infrastructure. This is an extension of the already popular hosted application model.

• *Platform as a Service (PaaS)*, where the product is a development and deployment platform, a set of APIs, libraries, programming languages and associated tools used for application creation. Users of PaaS are developers and companies which create

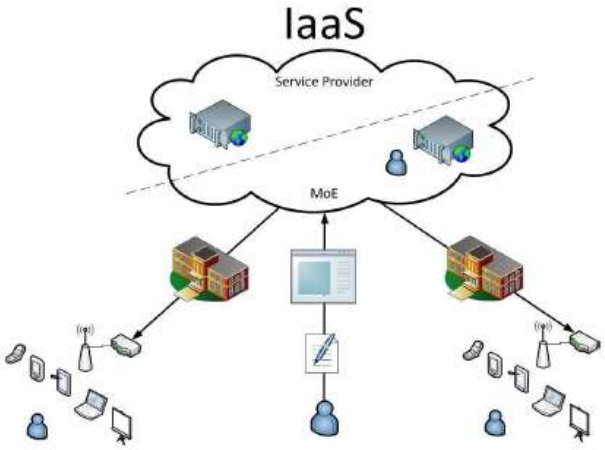
customized applications for end-users, and as such they are allowed to have control over some aspects of the application's environment but without direct access to the operating system and the hardware.



• *Infrastructure as a Service (IaaS)*, where the product is the low-level infrastructure used to create customized application environments or even higher level products (which might be PaaS or SaaS).

Users of IaaS are given complete control of their infrastructure resources (most notably, virtual machines) and can configure and use them as they see fit. IaaS is the lowest-level type of cloud service and it does not usually carry the obligation to use any prescribed technologies (though the hosted environments may be preconfigured).

These models progress from a high-level service model directly accessible to end-users to more low-level services in which their immediate users have control over the application or the basic infrastructure.



A. Cloud computing examples and target users Each of the models has its primary audience, its strengths and weaknesses. SaaS is already popular among endusers in the form of publicly available, widely used hosted applications like web-mail applications (e.g. Google Mail1, HotMail2, Yahoo Mail3), picture sharing applications (e.g. Picasa4, Flickr5), video clip sharing

applications (e.g. YouTube6, Vimeo7), and some forms of office applications (e.g. Google Apps8, Microsoft Exchange Hosted Services9, Microsoft Office Live10).

Such services are provided without giving users access to any advanced application, infrastructure or hardwarelevel configuration or management features. This benefits users who do not want to concern themselves with the technical aspects of the service, while allowing providers to reduce costs through mass deployments without significant reconfiguration and integration [6].

The PaaS model is oriented towards application developers and integrators, offering a common development and deployment platform for new applications or the customization of existing ones. It is successfully offered by Google (Google App Engine11),Microsoft (Windows Azure12) and Salesforce.com(Force.com13), among others. The model strongly focuses on developing applications that make use of the elasticity features offered by the platform, leaving lower-level tasks to the provider. As with the SaaS and the PaaS model, IaaS offers services to users, while removing a certain level of responsibility.

The users are given a larger degree of control over assigned resources, including storage, CPU and network resources, usually by allowing direct control of virtual machines. The properties of easy access, ondemand self-service and elasticity separate IaaS in cloud computing context from server hosting (and collocation) offered by a large number of companies. IaaS is implemented globally by providers such as Amazon(Amazon EC214), Rackspace (Rackspace Cloud15), FlexiAnt (FlexiScale16) and others. The IaaS model allows the greatest flexibility for users that can make use of it. It is the least complicated for providers, which need only concern themselves with the general infrastructure and running of the virtual machines, leaving users to manage the virtual machines' contents. Open source IaaS solutions suitable for enterprise use are the primary focus of this paper.

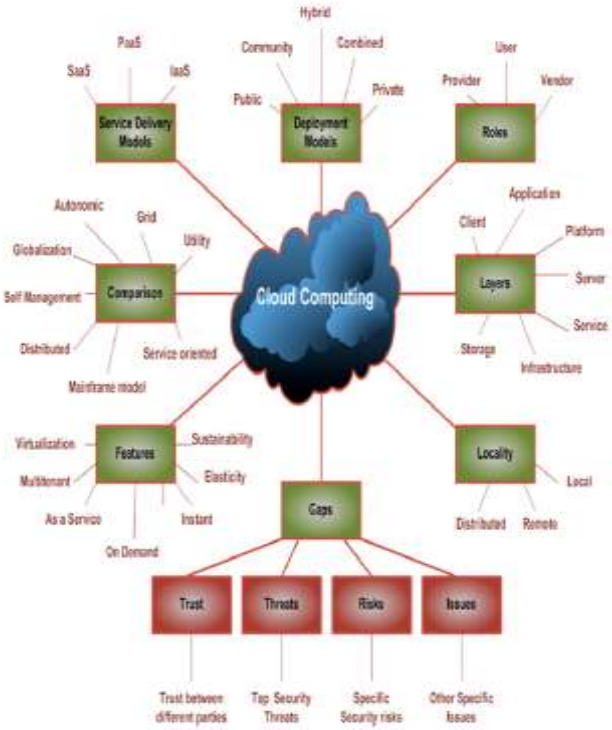


Fig. 1. Understanding cloud computing.[18]

3. *DEPLOYMENT MODELS*

The deployment models are orthogonal to the service models and describe the availability of the cloud deployments [3]. The three basic deployment models are:

• ***Private clouds***, used exclusively by one organization, usually operated internally by theorganization.

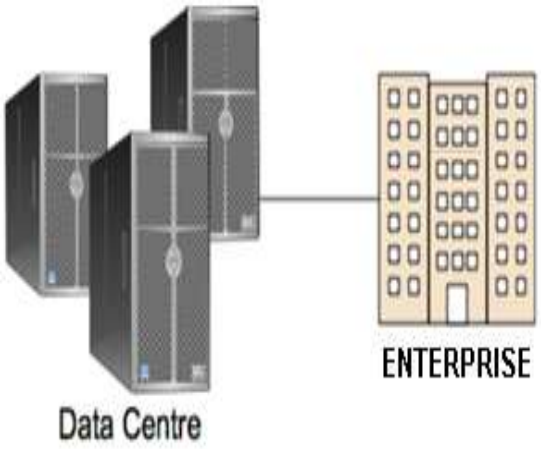
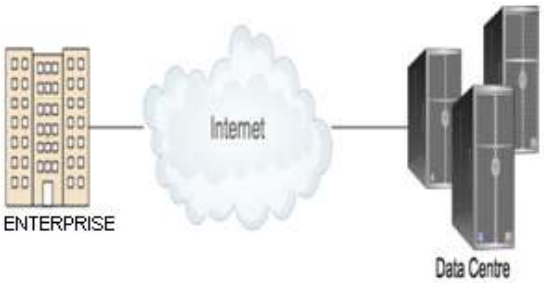


Fig2.privatecloud

• ***Public clouds***, used by generally interested parties(usually for a fee) in various ways. They can be



anextension of the "hosting provider" business model.

Fig3.public cloud

• ***Community clouds***, used by parties interested in specific requirements, such as organizations working on the same projects or on the sameproblem.

In addition to these three models, literature describes afourth model: a *hybrid cloud model* which combines two or more of the basic models. It is the most flexible model as it allows migrations of groups of resources between the categories.

In this work we focus on the *private cloud model* which can be used by enterprises to realize the benefits of cloud computing while still retaining control over the infrastructure they use [8][9].

4. Security Requirements In IaaS, PaaS, SaaS, [1]

There are already many existing laws and policies in place which disallow the sending of private data onto third-party systems. A Cloud Service Provider is another example of a third-party system, and organizations must apply the same rules in this case. It's already clear that organizations are concerned at the prospect of private data going to the Cloud. The Cloud Service Providers themselves recommend that if private data is sent onto their systems, it must be encrypted, removed, or redacted. Cloud computing is a service oriented Architecture which reduces information technology overhead for the end-user and provides great flexibility, reduced total cost of ownership ondemand services and many other benefits. Hence it delivers all IT related capabilities as services rather than product. Services on cloud are divided into three broad categories: software as a service, infrastructure as a service & platform as a service. Same as a service cloud is also classified as private cloud, public cloud & hybrid cloud.

Tabel 1. Security Requirements In IaaS, PaaS, SaaS

SaaS	PaaS	IaaS	Security Requirements
	✓	✓	Availability,resourcemanagement, trust, protection of communications, protection of the network and its resources, compliance, secure architecture, reliability and management of images
		✓	Control and governance, continuity of operations, risk management, protection of virtualization cloud, security hardware, hardware reliability, trusted third party, the basic configuration, the configuration change control, key management, connectivity to information systems, storage and computing
	✓		Identifying security threats, monitor configuration changes, procedures and security planning policy, accreditation, security, justice
✓		✓	Identity and Access Management
✓	✓		Anonymous

5. The Security Threats In The cloud layer

Because, the other two layer of service based on the cloud infrastructure layer are, security and management of cloud infrastructure services layer is very important security issues arise, storage facilities and processing services on a network as standard services We are, like server, switches, routers and so should be able to manage complex applications. other hand, cloud service providers to provide better resources to customers to use cloud computing services to make have.

The question then arises "How can the private data be automatically encrypted, removed, or redacted before sending it up to the Cloud Service Provider". It is known that encryption, in particular, is a CPU-intensive process which threatens to add significant latency to the

process. Large organizations using Cloud services face a dilemma. If they potentially have thousands of employees using Cloud services, must they create thousands of mirrored users on the Cloud platform? The ability to circumvent this requirement by providing single sign-on between on-premises systems and Cloud negates this requirement.

SaaS	PaaS	IaaS	Security threats	Row
×	√	√	Security threats	1
√	√	√	Insecure programming interfaces	2
×	×	√	Remove unsafe and incomplete data	3
×	×	√	Threats virtualization	4
√	√	√	Loss or data leakage	5
√	√	√	Hijacking Service	6
√	√	√	Personnel uncertain	7
√	√	√	Authorized change	8
√	√	√	Support research	9
√	√	√	Risk management interface	10
√	√	√	Traffic flow analysis	11
√	√	√	Connection failures and disruption of communication	12
×	×	√	Dependence secure hypervisor	13
×	×	√	Multitenant	14
√	√	√	Share issues related to technology and technology	15
√	√	√	Unknown risk profile	16
√	√	√	Attract hackers	17

Tabel 2.The Security Threats In The cloud layer [18]

6. OPEN SOURCE CLOUD COMPUTING PRODUCTS

We have selected a number of open source products which we consider to have a viable future for applications in enterprise environments. We intended to include the Enomaly ECP Community Edition17 but due to discontinuous work, we decided to omit it.

A. OpenNebula

OpenNebula18 is an open source software toolkit for cloud computing, which can be used to build and manage private, public and hybrid clouds. Since it does not contain virtualization, network, storage or security technologies, its primary use is as an orchestration tool for virtual infrastructure management in data-centers or clusters in private clouds and as merger of local and

public cloud infrastructure supporting hybrid scalable cloud environments.

Some of the main principles which guided the design of OpenNebula are full openness of architecture and interfaces, adaptability to various hardware and software combinations, interoperability, portability, integration, stability, scalability and standardization. Its main features include data-center or cluster management with Xen, KVM or VMware virtualization. It leverages the most common cloud interfaces Amazon AWS, OGF OCCI and VMware vCloud, and provides user management with authentication, multiple user rolling, secure multi-tenancy

and quota management. In the scope of cloud management a rich set of storage, virtual image, virtual machine and virtual network management features is provided. It supports cloud-bursting with Amazon EC2, simultaneous access to multiple clouds, and cloud federation. Standardization and interoperability are supported through abstraction from infrastructure and modular approach. Standard APIs includes Ruby, Java and XMLRPC. Security concerns are addressed with internal and external SSL communication and LDAP integration. OpenNebula EcoSystem adds a set of tools, extensions and plugins to OpenNebula Cloud Toolkit componentsenabling integration with existing products, services and management tools for virtualization, clouds and data centers. Telecom and hosting market,

and respectable scientific organizations like CERN adopted OpenNebula.

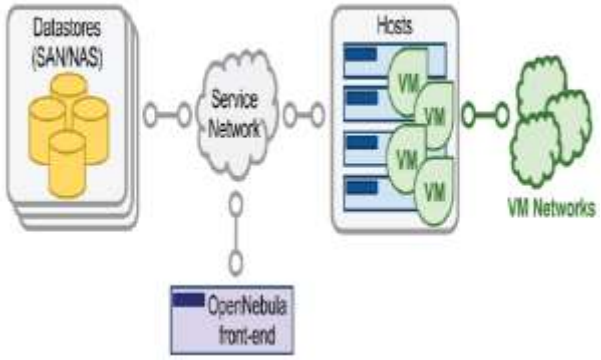


Fig4.Opennebula

B. Eucalyptus

As you can see, Eucalyptus composed of different elements, such as the cloud controller, cluster controller, storage controller, SC and NC have been formed.Cloud controller through a Web interface that provides the user the possibility to gain your virtual machine. Infrastructure also interacts with internal components. Storage controller acts as a storage system to protect files and virtual machine images. This component, such as a storage tank controller is permanent and stable. Information and samples of reservoir controller also makes maintenance and protection. Cloud controller and controller nodes associated with the cluster controller and task management and send commands to run on the controller node is responsible for the samples. Each cluster controller might consist of one or more controller nodes that all components monitor and manage them. In fact, the information received from the nodes to the cloud controller. Node controller, a machine that is installed Fvqnazr and do arithmetic operations. Also, the implementation, control and remove virtual machines on the Dard.kntrlknndh node by sending a query to the operating system, the amount of resources available, including the number of CPU cores, the RAM memory and disk space acquire this information to the cluster controller sends.Open source cloud computing architecture Eucalyptus[12] provides a scalable IaaS framework forimplementation of private and hybrid clouds. It wasinitially developed to support high performance computing (HPC) research at the University of California, Santa Barbara, and engineered to ensure compatibility with existing Linux-based data centers. It is component-based, flexible and highly modular with welldefined interfaces. Main design goals were simple

installation, non-intrusion and standardized language-independent communication. Eucalyptus also provides avirtual network overlay that isolates user network traffic and allows multiple clusters to appear as in the same LAN. Eucalyptus implements the Amazon Web Service (AWS) API allowing interoperability with existing services, enabling the possibility to combine resources from internal private clouds and from external public clouds to create hybrid clouds. This capability presents seamless integration with Amazon EC2 and S3 public cloud services. Eucalyptus currently supports Xen and KVM virtualizations, with plans to support others.

Four high level components are implemented as Web services. Cloud Controller (CLC) is a set of resource, data and interface services used for managing resources via node manager's queries, scheduling and cluster controller requests, visible as the main user interface. Storage Controller (Walrus) is a data storage service compatible with Amazon's S3 interface and Web services REST and SOAP interfaces. It accesses and stores virtual machine images and user data. Node Controller (NC) controls the execution, resources availability, and authorization on the host node. Cluster Controller (CC) collects information about a set of NCs, schedules run requests to NCs, and controls the instance virtual network overlay. Eucalyptus can be deployed on all major Linux OS distributions, including Ubuntu, Red Hat Enterprise Linux, CentOS, openSUSE, and Debian. Eucalyptus software core is included in Ubuntu distributions as a key component of the Ubuntu Enterprise Cloud.

According to [Nurmi et al 2009], the Eucalyptus project presents four characteristics that

differentiate it from others cloud computing solutions: a) Eucalyptus was designed to be simple without requiring dedicated resources; b) Eucalyptus was designed to encourage third-party extensions through modular software framework and language-agnostic communication mechanisms; c) Eucalyptus external interface is based on the Amazon API (Amazon EC2) and d) Eucalyptus provides a virtual network overlay that both isolates network traffic of different users and

allows clusters to appear to be part of the same local network.The Eucalyptus architecture is hierarchical and made up of four high level components, where each one is implemented as a stand-alone web service.

Node Controller (NC): this component runs on every node that is destined for hosting VM

instances. An NC is responsible to query and control the system software (operating system and hypervisor) and for conforming requests from its respective Cluster Controller. The role of NC queries is to collect essential information, such as the node's physical resources (e.g. the number of cores and the available disk space) and the state of VM instances on the nodes. NC sends this information to its Cluster Controller (CC). NC is also responsible for assisting CC to control VM instances on a node, verifying the authorization, confirming resources availability and executing the request with the hypervisor.

Cluster Controller (CC): this component generally executes on a cluster front¬end machine, or any machine that has network connectivity to two nodes:

one running NCs and another running the Cloud Controller (CLC). A CC is responsible to collect/report information about and schedule VM execution on specific NCs and to manage virtual instance network overlay.

Storage Controller (Walrus): this component is a data storage service that provides a mechanism for storing and accessing virtual machine images and user data. Walrus is based on web services technologies and compatible with Amazon's Simple Storage Service (S3) interface [Amazon 2006].

Cloud Controller (CLC): this component is the entry-point into the cloud for users. Its main goal is to offer and manage the Eucalyptus underlying virtualized resources. CLC is responsible for querying node managers for resources' information, making scheduling decisions, and implementing them by requests to CC. This component is composed by a set of web services which can be grouped into three categories, according their roles: resource services, data services, and interface services. While the details of the underlying resource architectures on which these systems operate are not commonly published, EUCALYPTUS is almost certainly shares some architectural features with these systems due to shared objectives and design goals. In addition to the commercial cloud computing offerings mentioned above (Amazon EC2/S3, Google AppEngine, Salesforce.com, etc.), which maintain a proprietary infrastructure with open interfaces, there are opensource projects aimed at resource provisioning with the help of virtualization.

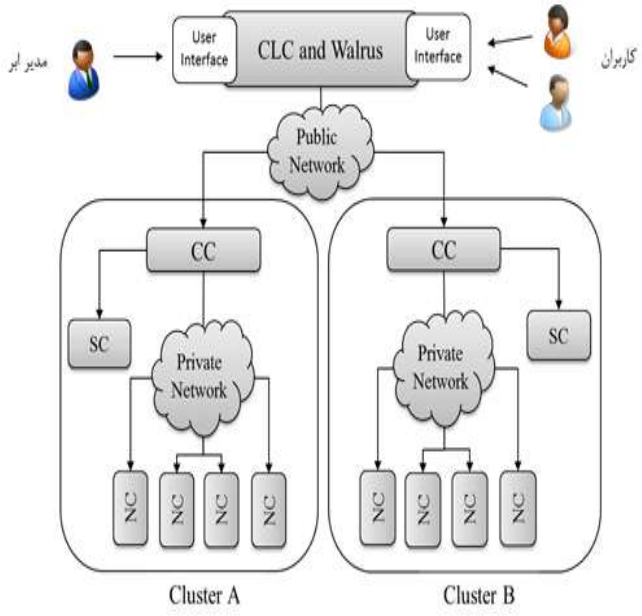


Fig5. Eucalyptus

Canonical's Ubuntu Linux distribution presents itself as a cloud OS with several cloud strategies, two of which are IaaS. Ubuntu Server Edition enables the use of Amazons EC2 but only as a public cloud. Ubuntu Enterprise Cloud (UEC)19 integrates Ubuntu Server Edition with

Eucalyptus over KVM hypervisor. The infrastructure of UEC is similar to Amazon's, but with simpler creation of private clouds.UEC [13] exposes five high-level components in a form of Web services: Cloud Controller (CLC), Walrus Storage Controller (WS3), Cluster Controller (CC), Node Controller (NC), and Elastic Block Storage Controller (EBS). The first four have the same functionalities as described in Eucalyptus overview, EBS runs besides CLC and provides persistent block devices and point-in-time volume snapshots stored on WS3. UEC defines security layers for authentication and authorization, network isolation, and machine instance isolation. Network isolation can be performed in four networking modes: system, static, managed and managed-noVLAN. Machine instance isolation is provided on three levels: networking, OS, and hypervisor-based machine.

D. OpenQRM

OpenQRM20 advertises itself as a data-center management platform. The core of openQRM follows the modern modular design and has no functionality itself, but instead focuses on abstracting storage and resources (computing resources, running virtual machines, VM images and other objects). OpenQRM features are provided via plugins which use the services exposed by the openQRM base. This architecture aims to make the whole system more stable and easier to manage as the base changes less often and provides a solid platform. OpenQRM can be installed on a variety of officially supported Linux operating systems: Debian, Ubuntu, SuSE, CentOS and Fedora. To achieve its goal of managed virtualized data-center, openQRM provides server and storage management, high-availability, realtime monitoring and virtual machine deployment and provisioning services, among others.

OpenQRM plugins provide a wide range of services, from integrated storage management (supporting directattached storage, and various SAN and NAS variants: iSCSI, LVM2, ATA-over-Ethernet and NFS), abstraction of virtualization (Xen, KVM, Linux-VServer, VMware Server and ESX VMs), migration from physical to virtual machines in three combinations (P2V, V2P and V2V of different VM type), high-availability (with failover from physical to virtual machines, and virtual to virtual failover between machines of same, or different type), and VM image templates or appliances.

E. Abiquo

Abiquo21 is a cloud management solution for virtualized environments in open source and commercial versions, mainly differing in resource limits, management and support options. Open source Abiquo Community Edition is licensed under LGPL Version 3. Main features include multi-tenancy, hierarchical user management and role based permissions with delegated control, resource limits, network, storage and workload management, multiple public, shared and private image libraries. It supports many Linux distributions (Red Hat, OpenSUSE, Ubuntu, Debian, CentOS, and Fedora), Oracle OpenSolaris, Microsoft Windows, and Mac OS X. Abiquo uses two storage systems: Appliances repository for virtual images in the form of NFS shared folder, and Virtual storage for virtual block devices available only in Enterprise edition. It distinguishes several types of server-side services: Java EE compatible application servers, database servers, cloud node servers, Appliance repository servers, and ISC DHCP servers. REST API can be used for integration with other systems. Abiquo server node incorporates Abiquo Core which contains the business logic, Appliance Manager for image library management and BPM that executes complex asynchronous tasks. Remote services deployed in the cloud expose system monitoring and management of virtual resources, physical machines, and storage.

Abiquo supports various virtualization technologiesincluding VMware ESX and ESXi, Hyper-V, VirtualBox,Xen, Citrix XenServer and KVM. Users of this solution benefit from powerful web management with functionalities such as drag-and-drop service deployment. It can be used for private clouds but also provides support for Amazon EC2. *F. Red Hat Cloud Foundations, Edition One* Red Hat22 offers a suite of open source software which provides infrastructure for public and private cloud solutions [11]. Red Hat Cloud Foundations, Edition One (RHCF) comprises of a set of products for virtualization, cloud, and application management and scheduling, but also operating systems, middleware, cookbooks, reference architectures with deployment instructions, consulting services, and training. RHCF Products are often tightly coupled with other Red Hat products. The suite comprises of Red Hat Enterprise Virtualization (RHEV), Red Hat Enterprise Linux (RHEL), Red Hat Network (RHN) Satellite, Red Hat Cluster Suite (RHCS), and Red Hat Enterprise MRG. RHEV for Servers is a product for end-to-end virtualization consisting of two components: RHEV Manager (RHEV-M) as a server virtualization system that provides advanced features (high availability, live migration, storage management, scheduler, etc.), and RHEV Hypervisor (RHEV-H), based on KVM hypervisor and deployed standalone or as RHEL hypervisor. RHN Satellite is a system management product providing software updates, configuration management, provisioning and monitoring across physical and virtual RHEL servers. RHCS is a clustering solution for RHEL supporting application/service failover and IP load balancing. Red Hat Enterprise MRG is a high-performance distributed computing platform providing messaging (MRG Messaging), real-time (MRG Realtime) and grid (MRG Grid) functionalities, and support for distributed tasks. Red Hat is investing and strongly participating in several cloud computing-related open source projects: Deltacloud, BoxGrinder, Cobbler, Condor, CoolingTower, Hail, Infinispan, Libvirt, Spice, and Thincrust. Red Hat also delivers JBoss Enterprise Middleware as a PaaS solution.

G. OpenStack

Collaborative software project OpenStack23, intends to produce an ubiquitous open source cloud computing platform that will meet the needs of public and private clouds regardless of size, at the same time be simple to implement and massively scalable.

Three interrelated components are currently under development: OpenStack Object Storage used for creation of redundant and scalable storage using clusters of commodity servers, OpenStack Imaging Service for retrieval of virtual machine images, and OpenStack Compute for provisioning and management of large groups of virtual private servers. OpenStack Compute represents cloud computing fabric controller and orchestrator for IaaS platform which can be used for management of various resources, networking, security, and access options. It defines drivers that interact with underlying virtualization mechanisms running on host and exposes functionality over a web-based API, but does not include any virtualization software. It is comparable to Amazon EC2 with additional support for projects that include volumes, instances, images, VLANs, keys and users. Images management relies on euca2ools (provided by Eucalyptus) and images are served through OpenStack Imaging Service or OpenStack Compute Service, supporting Amazon S3, OpenStack Object Storage or local storage. It also supports several virtualization standards including KVM, UML, XEN, Hyper-V and QEMU. OpenStack Compute can be deployed on Ubuntu, with tests on CentOS and RHEL under way.[23]

Fig6.Openstack

H. Nimbus

Nimbus24 is a set of open source software cloud computing components written in Java and Python targeting the needs of the scientific community, but also supporting other business use-cases. The main component is the Workspace service which represents a standalone site VM manager with different remote protocol frontends, currently supporting Nimbus WSRF frontend and partially Amazon EC2 with SOAP and REST interface. While Workspace service represents a compute cloud, there is also a quota-based storage cloud solution Cumulus, designed to address scalability and multiple storage cloud configurations. There are two types of clients: cloud clients for quick instance launch from various sites, and reference clients acting as full command-line WSRF frontend clients. Context Broker service allows clients to coordinate large virtual cluster launches using Context Agent, a lightweight agent on each VM. Context Broker manages a common cloud configuration in secure context across resources provisioned from potentially multiple clouds, with a possibility to scale hybrid clouds across multiple distributed providers. Nimbus supports the Xen or KVM hypervisors, and virtual machine schedulers Portable Batch System and Oracle Grid Engine The main advantage of Nimbus compared to OpenNebula is that it exposes EC2 and WSRF remote interfaces with attention to security issues, and can be combined with OpenNebula VM manager.

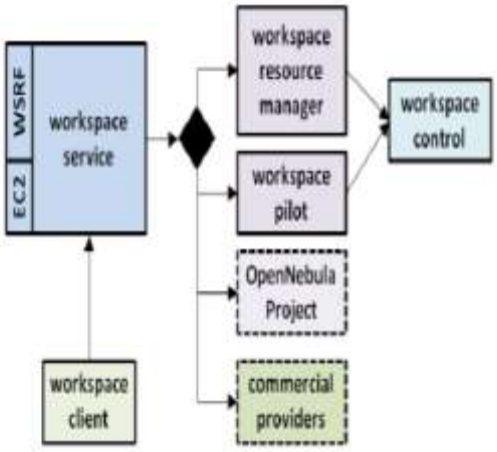


Fig7.Nimbus

I. mOSAIC

The main goal of the mOSAIC project [14] is the design of open source language- and platform independent API for resources and usage patterns that could be used in multiple cloud environments and construction of open source portable platform for cloud services. At the

current time there are no software deliverables, but the work is ongoing in cloud ontology, API description, testing environment, and usage patterns.

7. EVALUATION CRITERIA

Evaluating open source cloud computing products requires an elaborate set of evaluation criteria in order to provide a common baseline for IaaS cloud comparison. We have devised a set of 95 criteria which target features interesting for enterprise deployment. The criteria are grouped into six main categories: storage, virtualization, management, network, security and support. Storage-related criteria focus on supported approaches to storage: direct-attached storage, storage area network, and network-attached storage, as well as support for backup technologies and storage types. Virtualization criteria include virtualization types, support for actual virtualization technologies, and various monitoring and reconfiguration features, as well as support for migration and provisioning. Management features are essential for cloud implementers. The related criteria group captures features such as hardware and software integration, accounting, mass maintenance, reporting, and recovery.

Network features are highly dependent on the actual implementation, and the criteria focus on VLAN, firewall, performance, and integration support. Security criteria deal with permission granularity, integration with various directories, auditing, reporting of security events. Additional important features include storage encryption and secure management access.

OEM support is vital for enterprise deployment, and related criteria include an estimate of community vitality, vendor track record, possible support channels and SLAs, future viability of the product ecosystem, and completeness of provided free releases of the product.The criteria were devised with open source IaaS products in mind, but can be easily expanded to include commercial/closed technologies.

8. CONCLUSION AND FUTURE WORK

the solution presented And. these challenges, this area has become an important issue and the subject of security in cloud infrastructure services arise, the need for security among all levels of the cloud, most felt at the level of infrastructure as a service. becaus needs such as avaiability, reliability, data integrity, recovery, privacy and auditing in these areas become more tangible.

9. REFERENCES

[1] F. B. Shaikh, S. Haider, **"Security Threats in Cloud Computing"**, IEEE Internet technology and Secured Transactions (ICITST), 2011, pp 241-219.

[2] A.Bouayad et al, **"Cloud Computing: Security Challenges"**, IEEE Computer Knowledge and Technology 24, 2012, pp 26-31.

[3] I. Iankoulova, M. Daneva, **"Cloud Computing Security Requirements: a Systematic Review"**, Research Challenges in Information Science (RCIS) IEEE, 2012, pp 1 - 7.

[4] Dawoud, Wesam, Takouna, Ibrahim, Meinel, Christoph,**Infrastructure as a service security: Challenges and solutions**, Informatics and Systems (INFOS), 2010 The 7th International Conference onYear: 2010

[5] Djenna, Amir, Batouche, Mohamed,**Security problems in cloud infrastructure**,Networks, Computers and Communications, The 2014 International Symposium on year:2014,pp1-7.

[6] Hay, Brian, Nance, Kara, Bishop, Matt,**Storm Clouds Rising: Security Challenges for IaaS Cloud Computing**,System Sciences (HICSS), 2011 44th Hawaii International Conference on ,Year: 2011,PP:1-7,DOI: 10.1109/HICSS.2011.386.

[7] Chavan, Pragati, Patil, Premajyothi, Kulkarni, Gaurav, Sutar, R., Belsare, S,**IaaS Cloud Security**, Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on, Year: 2013 ,PP: 549 - 553, DOI: 10.1109/ICMIRA.2013.115

[8] Kumar, Saroj, Singh, Priya, Siddiqui, Shadab,**Cloud security based on IaaS model prospective**, Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference onYear: 2015,PP: 2173 – 2178.

[9] Winai Wongthai, Rocha, F., Van Moorsel, Aad,**Logging Solutions to Mitigate Risks Associated with Threats in Infrastructure as a Service Cloud**,Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference onYear: 2013 PP: 163 - 170, DOI: 10.1109/CLOUDCOM-ASIA.2013.70.

[10] Ristov, Sasko, Gusev, Marjan,**Security evaluation of open source clouds**,EUROCON, 2013 IEEEYear: 2013, PP: 73 - 80, DOI: 10.1109/EUROCON.2013.6624968.

[11] Litvinski, Oleg, Gherbi, Abdelouahed,**Openstack scheduler evaluation using design of experiment approach**,Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC), 2013 IEEE 16th International Symposium onYear: 2013 ,pp: 1 - 7, DOI: 10.1109/ISORC.2013.6913212.

[12] P. Sempolinski and D. Thain, **"A Comparison and Critique of Eucalyptus, OpenNebula and Nimbus"**, IEEE In Cloud Computing Technology and Science (CloudCom), 2010, pp 417-425.

[13] Donevski, A., Ristov, S., Gusev, M.,**Securityassessment of virtualmachines in opensourceclouds**, Information & Communication Technology Electronics & Microelectronics , (MIPRO), 2013 36th International Convention on, Year: 2013 PP: 1094 – 1099.

[14] Ristov, Sasko, Gusev, Marjan, Donevski, Aleksandar,**Security Vulnerability Assessment of OpenStack Cloud**,Computational Intelligence, Communication Systems and Networks (CICSyN), 2014 Sixth International Conference onYear: 2014 ,PP: 95 - 100, DOI: 10.1109/CICSyN.2014.32

[15] Haddad, Sammy, Dubus, Samuel, Hecker, Artur, Kanstrén, Teemu, Marquet, Bertrand, Savola, Reijo,**Operational security assurance evaluation in open infrastructures**,EUROCON, 2013 IEEE, PP:73 - 80 , DOI:10.1109/EUROCON.2013.6624968 .

[16] Achuthan, Krishnashree, SudhaRavi, Sreekutty, Kumar, Ravindra, Raman, Raghu, **Security vulnerabilities in open source projects: An India perspective**,Information and Communication Technology (ICoICT), 2014 2nd International Conference onYear: 2014 PP: 18 - 23, DOI: 10.1109/ICoICT.2014.6914033

[17] Bee Bee Chua, Bernardo, Danilo Valeros, **Open Source Developer Download Tiers: A Survival Framework**,IT Convergence and Security (ICITCS), 2013 International Conference on Year: 2013 PP: 1 - 5, DOI: 10.1109/ICITCS.2013.6717864

[18] Md.T.Khorshed, A.B.M.Shawkat, S. A.Wasimi,**"A Survey on Gaps, Threat Remediation Challenges and Some Thoughts for Proactive Attack Detection in Iass"**, Journal of Future Generation Computer Systems, 2012, 833–85

[19] M.Asadullah, R.K.Choudhary, **"Data Outsourcing Security Issues and Introduction of DOSaaS in Cloud Computing"**, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 18, January 2014, pp 40-44.

[20] N.Uma, **"Nelson, Semantic Based Resource Provisioning and Scheduling in Interlude Environment,"** Mobilizing resources in Latin America: The political economy of tax reform in Chile and Argentina. Palgrave Macmillan, 2012.

[21] K.Wood, M.Anderson, **"Understanding the Complexity Surrounding Multi tenancy in Cloud Computing",** IEEE Understanding the complexity surrounding multitenancy in cloud computing In e-Business Engineering (ICEBE), 2011, pp119-124.

[22] A. Abdullah, **"Resource Gate: A New Solution for Cloud Computing Resource Allocation"**, International Journal of Engineering and Technology Volume 2 No. 12, December, 2012

[23] http://www.OpenStack.org

[24] S. Ristov, et al, **"OpenStack Cloud Security Vulnerabilities from Inside and Outside"**, In CLOUD COMPUTING 2013, The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization, pp. 101-107. 2013

[25] http://archives.opennebula.org/documentation:rel4.4:plan

[26] Oliver Popović1et al, **"A Comparison and Security Analysis of the Cloud Computing Software Platforms"**, IEEE In Information Science and Engineering (ISISE), 2011, pp 632-634.

[27] Aleksandar Donevski, Sasko Ristov and Marjan Gusev, **Nessus or Metasploit: Security Assessment of OpenStack Cloud**,The 10th Conference for Informatics and Information Technology (CIIT 2013).

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 126 - 131, 2016, ISSN:- 2319–8656

Pricing Models for Cloud Computing Services, a Survey

Taj Eldin Suliman M. Ali
College of Graduate Studies,
Computer Science and Information Technology
Sudan University for science and technology
Khartoum, Sudan

Hany H. Ammar
Lane Department of Computer Science and Electrical Engineering,
College of Engineering and Mineral Resources
West Virginia University
Morgantown, USA

Abstract: Recently, citizens and companies can access utility computing services by using Cloud Computing. These services such as infrastructures, platforms and applications could be accessed on-demand whenever it is needed. In Cloud Computing, different types of resources would be required to provide services, but the demands such as requests rates and user's requirements of these services and the cost of the required resources are continuously varying. Therefore, Service Level Agreements would be needed to guarantee the service's prices and the offered Quality of Services which are always dependable and interrelated to guarantee revenues maximization for cloud providers as well as improve customers' satisfaction level. Cloud consumers are always searching for a cloud provider who provides good service with the least price, so Cloud provider should use advanced technologies and frameworks to increase QoS, and decrease cost. This paper provides a survey on cloud pricing models and analyzes the recent and relevant research in this field.

Keywords: Cloud Computing; Software-as-a-Service (SaaS); Service Level Agreement (SLA); Dynamic pricing; Quality of Service (QoS); revenue maximization; CSL.

1. INTRODUCTION

Cloud computing is an emerging parallel and distributed computing paradigm that depends on the internet to deploy computer resources and services dynamically and enables a provider to deploy a single application to execute in multiple machines based on a contract between the cloud providers and the consumers called Service Level Agreement (SLA). Traditionally, the resource/service price is defined in SLA and remains static. This static pricing mechanism has several problems such as overprovisioning as well as under provisioning problems. On the other hand, dynamic pricing is needed to overcome these problems. The central objectives of cloud provider are profit maximization and increase customer satisfaction level (CSL) i.e. the market sharing maximization. To achieve these objectives, the cloud provider need to reduce cost, SLA violations, response time, and power consumption; and deploy services in different prices (Dynamic pricing) based on the current consumer's requirements as well as the level of the offered QoS. On the contrary, the main objectives of cloud consumers are minimizing cost (price) and access services with high quality of services (QoS). To achieve all of these objectives - provider's and consumer's objectives -, the negotiation between them would be established through service level agreement (SLA).

"Service Level Agreement (SLA) is an agreement used to guarantee web service delivery, it defines the understanding and expectations from a service provider and service consumer" [1]. SLA is a legal contract that grantees Quality of Service (QoS) between the cloud provider and the consumers. This contract includes and defines many things, such as parties, services, prices, service level objectives (SLOs), obligations, penalties.

The business organization uses SLA to enlarge market sharing because, through SLA, the provider can increase CSL; as a result improve its profits. When SLA is violated, the CSL go down and some penalties would be enforced.

1.1 Background

The Cloud model is cost-effective i.e. the price is reasonable because cloud consumers only pay for their actual usage, they do not need to pay any upfront costs. Also, it is elastic as the cloud provider can deliver more or less according to the customers' needs.

Cloud provider offers different services to cloud consumers [14] at different prices, therefore, two stakeholders – cloud provider and cloud consumer - would be communicated and negotiated about several things such as QoS, price, etc. All of the negotiation points would be written in SLA clearly. Pricing represents an important indicator for success business companies which provide services or products [15]. Cloud provider uses several pricing models to specify the price. This pricing model would be established properly to define the fair price for both stakeholders (providers and consumers). A good pricing model supports cloud providers to achieve their objectives such as profit maximization; meanwhile it considers the cloud consumers.

1.2 Cloud Computing

1.2.1. Cloud Computing Definitions

Definition 1: *"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction." [17].*

Definition 2: *"A Cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumer." [4].*

Definition 3: *"Cloud computing is a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services*

are delivered on demand to external customers over the Internet."[3]

1.2.2. Cloud Computing Deployment Models

Cloud Computing approach is designed mainly to achieve many services generally and services deployment particularly. To deploy services, Cloud computing has several deployment models, such as public, private, community, and hybrid cloud. The brief descriptions for these models as follows:

- **Public Cloud:** through this model, services could be provisioned for general consumers, and cloud provider responsible for several thing main of them is infrastructure.
- **Private Cloud:** it is a model by which services could be deployed for particular consumers in an organization, the needed infrastructure owned and managed by the organization or by other third party organization.
- **Community Cloud:** at this model, cloud provider deploys services to organizations that have common objectives and policies, the needed infrastructure could be managed by one of these organizations or by the third party.
- **Hybrid Cloud:** Sometimes environment requiring a combination of two or more mentioned deployment models, at this case service provisioned by other model called a *hybrid cloud*.

1.2.3. Cloud Computing Service Models

Cloud computing providers offer their services for cloud consumers through using three fundamental models [5], software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS).

- **SaaS model:** at this model, consumers use the internet to access services (applications) which could be hosted in service provider's company. There would be two cornerstones, the provider, and the consumer. Service provider controls everything whereas service consumer controls application settings only. The common examples of SaaS are Facebook and Twitter. [5]
- **PaaS model**: through this model, providers deploy several things such as tools, programming environments, and configuration management for consumers to support them to deploy and develop the application. PaaS consumer's examples are software designers, testers, administrators, and developers. Common example of PaaS is Google AppEngine [8].
- **IaaS model**: this model enables providers to deploy virtual machines (VMs); storage places …etc for service consumers to help them to build systems. Some example of IaaS consumers are system administrators and system developers. Common example of IaaS is Amazon EC2 [6].

1.3 Objective

The objective of this paper is to survey current research on cloud computing pricing models, and analyze and compare their characteristics. We focus on SaaS services. We compare static and dynamic models, and identify the weaknesses of existing models.

1.4 Organization

The rest of the paper organized as follows: in the next section the researcher introduces Cloud Computing pricing models, its important concepts, classification, and provides some examples. In section three the paper focuses on the related works to cloud pricing, discusses some of the existing works and draws a comparison among the common existing pricing models. In the last section, the paper gives relevant conclusion and offers some suggestions.

2. CLOUD COMPUTING PRICING MODELS:

Demands and revenues are controlled by several factors. And "pricing" is considered to be the most important one. Cloud providers always use pricing to know (1) how much service provisioning could be done for different consumers (2) the relationship between pricing and other issues such as provisioning period and grant discounts.

2.1. Cloud Computing Pricing Model classification:

As mentioned in [7, 21], the two common types of pricing models are:

- **Fixed Pricing Model:** Here the price charging doesn't change, and the cloud provider is someone who determines the price to the resource type in advance. For example, Amazon provides disk space for $0.15/GB, and service consumers have the same services at all time, such as Pay-per-use model. According to Yeoa et al. [2], fixed pricing model is more straightforward and easy to understand, but it is unfair for all customers because they are not having the same needs.
- **Dynamic Pricing Model:** In this model the price charging changes dynamically according to market status quo. The service price could be calculated for each request according to the pricing mechanism that is used. In this case, service consumer requests and receives several types and levels of services in need, such as Market-dependant pricing model.

Table1 bellow shows the strength and weakness of the above types of pricing models:

Table1. Fixed pricing vs. dynamic pricing

Pricing Model	Advantages	Disadvantages
Fixed pricing model	• It supports assurances for consumers. • Consumers know how much they will pay. • More consistent. • It reduces risks. • Make profit estimation easy.	• Unfair for consumer: If the user doesn't consume the resource extensively, he/she may pay more than his/her real utilization. • It does not allow provider to change price at any account. • Unfair for provider: During proper resource utilization consumer may pay less than his/her real utilization.
Dynamic pricing model	• It supports provider to maximize profits with each consumer.	• Some consumers are not interested in this model as they prefer a fixed price to dynamic price. • Consumers who pay more feel inequality

	• Fair for consumer as it enables him to pay according to the offered QoS. • It supports provider to set price based on current state of the market (season or supply and demand)	consequently having negative opinions. • In some environments such as entertainment sites consumers do not prefer dynamic pricing.

Financially, the economic efficiency could be measured by two indicators i) the number of allocated resources for providers, and the total of the achieved consumers' requests. ii) The average of the consumers' welfare. Authors of [20] draw a comparison among the number of the achieved consumers' requests, the number of provider's allocated resources, and the average of consumer's welfare with a fixed pricing.

According to [9], in cloud computing, there are several points that determine the price, as follow:

- **The annual costs:** it is the fees that cloud provider pays annually to buy the resources.
- **The period:** it is the leasing period of resources by consumers. In this respect some provider confirms that service price could be decreased if the leasing period is long.
- **Quality of Services:** it is the quality assurance that the cloud providers use to identify the entity of service. Such as service availability, security, and privacy. Some providers increase service price if the level of QoS for service is high.
- **The level of resource:** it is the age of resource that the cloud consumer rents. Some providers say that older resources mean lower price.
- **Maintenance fees:** it is the annual cost to maintain and secure resources. Some rented resources become weak or damaged because of the continuous use. These resources could be maintained.

Consumers can assess the providers depending on the following factors:

- **Pricing Scheme:** the mechanism in which service price could be determined, such as pay-as-you-go model.
- **Service Customizability:** the way by which provider customizes his SaaS services to meet service consumer's requirements.
- **Leasing period:** the period, at which customers can consume services, some examples of this period are subscription, pay-per-use, and perpetual.
- **Service QoS:** the mechanism by which service requirements could be specified; such as scalability, availability, and security.

2.2. Examples of Cloud Pricing Models:

There are several pricing models, as follow:

- **Pay-as-you-go Model:** in this model, the price could be determined by the cloud provider and remains static. Customer pays a fixed price and reserves resources due to the paid period. On the contrary, the customer may utilize the resource improperly i.e. the consumer gets less benefit.[8]
- **Subscription Model:** here, the price depends on subscription period. If the consumer utilizes resources extensively, the underprovisioning problem will occur. Also, the overprovisioning problem occurs when the customer is not consuming the resources extensively. [8]
- **Pricing algorithm for cloud computing Resources:** this is a real-time approach, in which provider can maximize revenues and minimize cost. This approach is just a theoretical, and not yet applied. [12]
- **Dynamic resource pricing on federated clouds:** it is a fair theoretical approach for both provider and consumers because it is dynamic – the price depends on the level of supply and demand. [10]
- **Competition-based pricing model:** it is a dynamic approach because the determined price depends on competition. This approach could be implemented easily, but it neglects the customers. [13]

3. RELATED WORK

This section discusses several existing works, presents weak points, and finally compares some of the pricing models.

In [9] the authors proposed a *novel financial economic model*, in which customers can gain a high level of QoS. The authors noted that the optimal price by which service provider can recover the initial cost, which was defined between two boundaries. They use the financial option theory to define the lower boundary and the Moor's law to define the upper boundary. However, it disregards the maintenance costs. In [6, 8, 11] the authors implemented *pay-as-you-go pricing model*, by which service provider can determine a fixed price. Here, if there is a high demand, the service provider is not allowed to change the period of (resource reservation) or raise a price. On the other side if the demand is low, the consumer negatively pays more than his/her real usage. In [6, 8, 11], the authors implemented *subscription pricing model*, by which the price determined according to a period of subscription. This model is good if the customer consumes the service extensively. However, a consumer may do not consume service properly (pay more than use).

In [12], authors introduced *Pricing algorithm for cloud computing resources*, that could be used for minimizing cost as well as maximizing profits for the service provider. However, this is a fixed model and not suitable on supply/demand changes.

Dynamic resource pricing on federated clouds was introduced in [10], by which the price could be determined depending on the level of supply and demand. However, this model does not support a good scalability during high demand period.

In [13], authors implemented *Competition-based pricing model*, by which provider sets the price according to competitors. However, in this model, consumers are not taken into consideration.

Customer-based pricing model in [2] was introduced that; the price could be specified according to the customers' needs (what the consumer ready to pay). However, the consumer does not know what he/she is ready to pay at every time.

Table2 shows a comparison among several cloud pricing models, considering the following criteria: the mechanism to determine the price, whether the model is static or dynamic, and the advantages and disadvantages.

Table2. Comparison of several cloud pricing models

#	Pricing Model	Type (Static/ Dynamic)	Nature (Implemented/ theoretical)	Mechanism	Advantages	Disadvantages
1	Subscription Model [8, 11]	Static	Implemented	Cloud provider defines Resource/Service prices depending on lease period	It is good for consumer when Resources/Services are utilized extensively	Consumer may pay more than the real utilization cost when he/she does not use Resources/Services properly
2	Pay-as-you-go Model [8, 11]	Static	Implemented	Cloud provider determines a constant Resource/Service price	Resources/Services are available during reservation period, and the price is known	Overprovisioning and underprovisioning problems may occur. The price is unchangeable
3	Pay-for-resources model [8, 11]	Static	Implemented	Cloud provider determines Resource/Service prices according to the cost.	Maximizes resource utilization	Difficult to be implemented
4	Dynamic resource pricing on federated clouds [12]	Dynamic	Theoretical	Cloud provider uses current level of supply/ demand to determine Resource/Service prices	Increases consumers' satisfaction and maximizes the number of their profitable requests	It does not support a good scalability during high demand period
5	Value-based pricing [20]	Dynamic	Implemented	Resource/ Service prices are defined depending on the customer's point of view	Increases revenues	Hard to implement
6	Competition-based pricing [13]	Dynamic	Implemented	Cloud provider uses competitors' prices to determine the current price for service/resource	Easy to implement	Ignores the cloud customers
7	Datacenter net profit optimization with individual job deadlines [18]	Dynamic	Theoretical	Cloud provider uses job scheduling mechanisms to set Resource/ Service prices	Maximizes cloud provider's revenues, minimizes power consumption cost	It doesn't take in consideration the heterogeneous servers. Difficult to implement
8	Genetic model for pricing in cloud computing markets [17]	Dynamic	Theoretical	Price is specified by cloud provider depending on the state of a real time market.	Maximizes revenues, flexible implementation	Very critical during the (rise and fall) demand period.
9	A novel financial economic model [9]	Dynamic	Theoretical	Cloud provider sets Resource/Service prices between upper and lower boundaries	Maximizes profits for cloud provider and improves QoS for cloud consumer	Maintenance costs are not taken in consideration.
10	Customer-based pricing [13]	Dynamic	Implemented	Cloud consumers define the current price	Cloud consumer is taken into consideration	Difficult to set price
11	Cost-based pricing [19]	Dynamic	Implemented	Cloud provider specifies profit level to set Resource/Service prices	Cloud provider can define the price easily	It doesn't considers cloud consumer
12	Pricing algorithm for cloud computing Resources [10]	Dynamic	Theoretical	Resource/Service prices are set according to the current market state.	It is better for cloud provider because it maximizes revenues by reducing cost	Useless when supply/demand differ quickly

4. CONCLUSION

In this paper, we surveyed different types of cloud pricing models. We compared static pricing model versus dynamic pricing model. Based on this comparison, we conclude that on one hand the static model is easy for both understand-ability and profit estimation but some problems such as under provisioning and over provisioning may occur. On the other hand the dynamic pricing model is fair for consumers because it supports them to pay depending on the QoS required; also it is fair for the provider so it help him to maximize profits.

Also, during this survey, we presented detailed comparison among twelve pricing models based on the following factors: the type (static/dynamic), nature (theoretical/implemented), the mechanism to determine price, the advantages, and disadvantages. Depending on this comparison we note that all of the static models are implemented but some of the dynamic models are theoretical, on the static models the provider defines the price but on the dynamic models the price could be defined by the provider to maximize revenues and rarely optimized for the consumers.

In summary, due to the fact that cloud computing services have dynamic behavior with pay-as-you-go models, it is necessary to conclude that the dynamic pricing models are much more adequate for the consumers because they adapt to different variable needs. Also, they are batter for the providers because they need to support Multi-Tenants and change (increase/decrease) in the price depending on the market state. Finally, we noted that most of pricing models favor the providers over the consumers. Our research suggests that there is a need for new pricing models that take the two stakeholders, the provider and the consumer, in its consideration.

5. REFERENCES

[1] Jin, L. J., and Machiraju, V. A. Analysis on Service Level Agreement of Web Services, (June 2002).

[2] C. S. Yeoa, S. Venugopalb, X. Chua and R. Buyyaa, “Autonomic Metered Pricing for a Utility Computing Service”, Future Generation Computer Syst., vol. 26, no. 8, (2010).

[3] Foster I, Yong Z, Raicu I, Lu S Cloud computing and grid computing 360-degree compared. In: Proc 2008 grid computing environments workshop, pp 1–10, (2008).

[4] Buyya, R., and Alexida. D. A Case for Economy Grid Architecture for Service Oriented Grid Computing. In Proceedings of the 10th International Heterogeneous Computing Workshop (HCW), San Francisco, CA, (2001).

[5] W. Voorsluys, J. Broberg, and R. Buyya, Introduction to Cloud Computing,” Cloud Computing: Principles and Paradigms, chapter 1, pp. 1–41.Technical Report HPL-2002-180, Software Technology Laboratories, HP Laboratories, (2011).

[6] Varia, J. Architecting Applications for the Amazon Cloud. Cloud Computing: Principles and Paradigms, Buyya, R., Broberg, J., Goscinski, A. (eds), ISBN-13: 978-0470887998, Wiley Press, New York, USA. Web - http://aws.amazon.com. (2010).

[7] A. Osterwalder, “The Business Model Ontology – A Proposition in a Design Science Approach”, Doctoral thesis, University of Lausanne, (2004).

[8] Google App Engine, https://appengine.google.com/.(2015)

[9] B. Sharma, R. K. Thulasiram, P. Thulasiraman, S. K. Garg and R. Buyya, “Pricing Cloud Compute Commodities: A Novel Financial Economic Model”, Proc. of IEEE/ACM Int. Symp. on Cluster, Cloud and Grid Computing, (2012).

[10] M. Mihailescu and Y. M. Teo, “Dynamic Resource Pricing on Federated Clouds”, Proc. 10th IEEE/ACM Int. Symp. on Cluster. Cloud and Grid Computing, (2010).

[11] Windows Azure, http://www.windowsazure.com/en-us/.

[12] H. Li, J. Liu and G. Tang, “A Pricing Algorithm for Cloud Computing Resources”, Proc. Int. Conference on Network Computing and Inform. Security, (2011).

[13] J. Rohitratana and J. Altmann, “Agent-Based Simulations of the Software Market under Different Pricing Schemes for Software-as-a-Service and Perpetual Software”, Economics of Grids, Clouds, Systems, and Services, ser. Lecture Notes in Computer Science, Altmann et al., Eds. Springer Berlin/Heidelberg, pp. 6296. (2010).

[14] A. Monaco, “A View inside the Cloud”, http://theinstitute.ieee.org/technology-focus/technology-topic/a-view-inside-the-cloud.

[15] S. Dutta, M. Zbaracki and M. Bergen, “Pricing Process as a Capability: A Resource-Based Perspective”, Strategic Management Journal, vol. 27, no. 7, (2003).

[16] M. Macias and J. Guitart, “A Genetic Model for Pricing in Cloud Computing Markets”, Proc. 26th Symp. of Applied Computing, (2011).

[17] W. Wang, P. Zhang, T. Lan and V. Aggarwal, “Datacenter Net Profit Optimization with Individual Job Deadlines”, Proc. Conference on Inform. Sciences and Systems, (2012).

[18] S. Lehmann and P. Buxmann, “Pricing Strategies of Software Vendors”, Business and Information Systems Engineering, (2009).

[19] P. Nähring, “Value-Based Pricing”, Bachelor Thesis, Linnaeus University, (2011).

[20] M. Mihailescu and Y. M. Teo, “On economic and computational-efficient resource pricing in large distributed systems,” in Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on, may 2010, pp. 838 –843.

[21] Samimi, P.; Patel, A.; , "Review of pricing models for grid & cloud computing," Computers & Informatics (ISCI), 2011 IEEE Symposium on , vol., no., pp.634-639, 20-23 (March 2011).

6. AUTHRORS BIOGRAPHIES

Taj Eldin Suliman M. Ali: B.Sc. in Computer science at Sudan University (SUST), and M.Sc.in Computer science at Khartoum University. From 2009 to 2015 he was work as a lecturer and academic coordinator, College of Computer Studies at National Ribat University, now he is a lecturer. From 2010 to 2015 he was work as a lecturer and academic coordinator, College of Computing and Health informatics at National University – SUDAN, now he is a lecturer. During 2006 to 2008 he was work as a lecturer at Bayan Collage, 2008 to 2009 he was work as Dean of IT department at Bayan Collage. From 2002 to 2005 he was work as TA at Bayan Collage and also works as a programmer. Currently, he is a Ph.D. student at Sudan University of Science and Technology, College of Computer Science and Information Technology (SUST), my research interests in Pricing models and Cloud Computing.

Hany H. Ammar BSEE, BS Physics, MSEE and Ph.D. EE, is a Professor of Computer Engineering in the Lane Computer Science and Electrical Engineering department at West Virginia University. He has published over 170 articles in prestigious international journals and conference proceedings. He is currently the Editor in Chief of the Communications of the Arab Computer Society On-Line Magazine. He is serving and has served as the Lead Principal Investigator in the projects funded by the Qatar National Research Fund under the National Priorities Research Program. In 2010, he was awarded a Fulbright Specialist Scholar Award in Information Technology funded by the US State Department - Bureau of

Education and Cultural Affairs. He has been the Principal Investigator on a number of research projects on Software Risk Assessment and Software Architecture Metrics funded by NASA and NSF, and projects on Automated Identification Systems funded by NIJ and NSF. He has been teaching in the areas of Software Engineering and Computer Architecture since 1987.In 2004, he co-authored a book entitled Pattern-Oriented Analysis and Design: Composing Patterns to Design Software Systems, Addison-Wesley. In 2006, he co-authored a book entitled Software Engineering: Technical, Organizational and Economic Aspects, an Arabic Textbook.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 132 - 136, 2016, ISSN:- 2319–8656

Identification of Spam Emails from Valid Emails by Using Voting

Hamoon Takhmiri
Computer Science and Technology
Islamic Azad University Kish International Branch
Kish Island, Iran

Ali Haroonabadi
Islamic Azad University Kish International Branch
Kish Island, Iran

Abstract: In recent years, the increasing use of e-mails has led to the emergence and increase of problems caused by mass unwanted messages which are commonly known as spam. In this study, by using decision trees, support vector machine, Naïve Bayes theorem and voting algorithm, a new version for identifying and classifying spams is provided. In order to verify the proposed method, a set of a mails are chosen to get tested. First three algorithms try to detect spams, and then by using voting method, spams are identified. The advantage of this method is utilizing a combination of three algorithms at the same time: decision tree, support vector machine and Naïve Bayes method. During the evaluation of this method, a data set is analyzed by Weka software. Charts prepared in spam detection indicate improved accuracy compared to the previous methods.

Keywords: Spam emails; tree decision; Naïve Bayes; support vector machine; voting algorithm

1. INTRODUCTION

In recent years, the increasing use of e-mails has led to the emergence and increase of the problems caused by unwanted bulk email messages commonly known as spam. By changing the common and aggressive content of some of these messages from a minor harassment to a major concern, spams began to reduce the reliability of emails. Personal users and companies that are affected by spams because of network bandwidth spent a lot of time for receiving these messages and distinguishing spam messages from standard messages (legally certified) for users. A business model based on spams for buying and selling is usually beneficial because the costs for the sender is little, so a lot of messages can be sent by maximizing replies and this aggressive performance is a feature of known spammers. Economic functions of spams have forced some countries to legislate for them. In addition, problems of pursuing the transmitters of these messages can limit the performance of such laws. In addition to legislation, some institutions have imposed changes in protocols and practical models. Another approach being implemented is using spam classifiers which by analyzing message content and additional information try to identify spam messages. For using this function once, messages are identified usually based on the settings used by classifier [1]. If classifiers are used by a single user, as a customer-focused classifier, messages are usually sent to a folder that only contain messages under the title of spam and this makes it easier to identify the messages. On the other hand, if the classifier works on the Email server, by checking multiple users' messages, they can be tagged as spam or get deleted. Another possibility is a multi-user setting in which classifiers running on different machines share information about the received messages to improve their performance. Generally, the use of classifiers has created an evolutionary scenario in which spammers use instruments with different ways, in particular, regulated methods for reducing the number of messages identified. Initially, spam classifiers were based on the user's known laws and were designed on the basis of arrangements that are easily seen in such messages. [2].

2. RELATED WORK

Filters have been relied on key word patterns. In order to be efficient and avoid the risk of accidental deletion, non-spam messages are called as legitimate messages or Ham. These patterns should be checked manually by each user's email. However, fine adjustment of patterns requires time and expertise which is unfortunately not always available. Even messages' features change over time which requires key word patterns to be updated regularly [3]. Thus, processing messages and detecting spams or non-spams automatically is desirable. It should be noted that text categorization methods can be effective in anti-spam filtering. Sending bulk unsolicited message makes it a spam message, not its real content. In fact, posting a bulk message carelessly makes the message a spam. Phenomena can be images, sounds, or any other data, but important thing is to distinguish between different samples and have a good reaction for every sample. Learning is usually used in one of the following ways: Statistical, synthetic or neural [4].

Recognition of Statistical pattern by assuming that these patterns are created by a possible system is determined based on the statistical properties of patterns. Some of important reasons for spamming include economic purposes, as well as promoting a product, service or a particular idea, tricking users to use their confidential information, transmission of malicious software to the user's computer, creating a temporary email server crash, generating traffic and broadcasting immoral contents. Spams are constantly changing their content and form to avoid detection by anti-spams. Some ways to prevent spamming include:

• Economic methods: getting cash to send e-mail: Zmail Protocols

• Legislative procedures: such as CAN-SPAM laws, securing email transmission platform

• Changing the e-mail transmission protocol and providing alternative protocols such as sending id and features

• Controlling your outgoing emails versus controlling incoming mails.

• filtering based on learning (statistical) and by using the mail features

• Mail detection phishing (fake pages) to help fuzzy classification methods

• Controlling methods: Controlling the features of a mail before sending it by e-mail server

3. SUGGESTED METHOD

For better identification of spams, the aim is initially finding behavioral features of spam, so the data mining and logging of spam behavior is required at first, such as detecting the IP of sender, sending time, frequency, number of attachments and so on. This information is stored in the database so that they are structure information. Behavioral attributes of spams can be extracted from these reports created in the mail server [5].

Before extracting data, analyzing the features of e-mails from reports is required. Information obtaining technology is selected to analyze these features and the main feature is obtained. Some features of fewer information and weaker relations are deleted. The behavioral feature of an individual e-mail is as follows:

- Customer IP (CIP)
- Receive Time (RT)
- Context Length (CL)
- Frequency (FRQ)
- Context Type (CT)
- Protocol Validation (PV)
- Receiver Number (RN)
- Attachment Number (AN)
- Server IP (SIP)

There are no entirely clear features in real world, so that after fuzzification, normal and logical degrees below horizon can be explained for characterization of samples. After preprocessing, the value of information is as follows:

A) Customer IP: is used only to calculate the frequency of sender and extracting the common behavioral pattern of sender and is not involved in the decision tree computing.

B) Receive Time: the value of day and night time is a common value and requires fuzzification for horizontal degree (0, 1).

C) Context Length: the value of short and long for the size of email is also a common value feature and requires fuzzification too.

D) Protocol Validation: is Boolean and when it complies with the sender is (1) and in case of non-compliance is (0).

E) Context Type (CT): The value a text or Multipart is (1) for the text and (0) for the Multipart.

F) Receiver Number (RN): the value of many and less is a common value feature and also requires fuzzification.

G) Frequency (FRQ): the value of often and seldom frequency is a common value feature and also requires fuzzification.

H) Attachment Number (AN): the value of many and less is a common value feature and also requires fuzzification. Table 1 lists some examples of the results after preprocessing.

Table 1 Result Of Data Processing

CIP	RT	CL	FRQ	CT	RN	PV	AN	SIP
...	...	...	...	...	...	...	...	...
IP1	15	4987	2	text	3	valid	0	SIP 1
IP2	15	890	4	html	1	valid	1	SIP 2
IP3	15	1298	1	html	1	invalid	0	SIP 1
IP4	16	2442	3	multipart	2	valid	2	SIP 3
...	.	...	...	...	...	...	...	...

It is assumed that (A, B) are fuzzy subsets defined in a confined space (F). “If A so B” is named as a fuzzy rule and simply recorded as (A → B), which is called fuzzy sets of conditions, and (B) is called fuzzy sets of conclusion. Based on the knowledge of decision tree, rules were classified as fuzzy and are in the form of "if - then". A rule is created for each path from the root to the leaves. Simultaneously with a special path, any value features are as a pair of (And) piece of a rule that is called prior rule. (Section If) predicts leaf node classification, so forms the compliance rule. (Section Then) of "if - then" rule are easier to understand, especially when the tree is large [6].

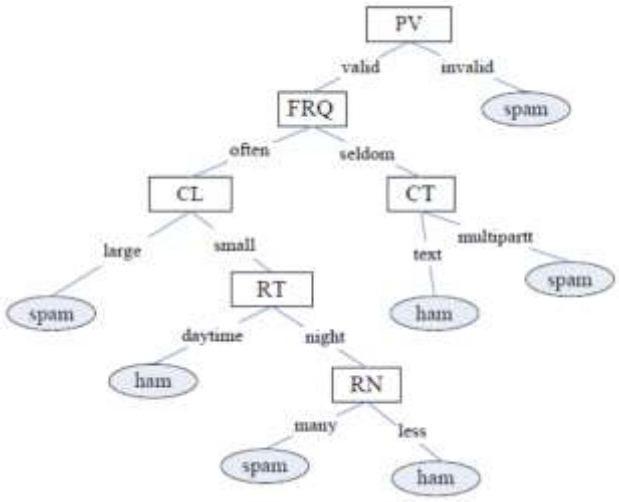


Figure 1 Decision tree

After checking decision trees and identifying important characteristics of a mail from the proposed decision tree in Figure 1 decision tree, Mamdani decision tree rules are generated as follows:

1. If the protocol (PV) of sending email is not authentic, the email is spam.

2. If e-mail protocol (PV) is valid, frequency (FRQ) is less and type (CT) is Multipart, the e-mail is spam.

3. If e-mail protocol (PV) is valid, frequency (FRQ) is many and context length (CL) is long, the e-mail is spam.

4. If e-mail protocol (PV) is valid, frequency (FRQ) is many, context length (CL) is short and receiver number (RN) is many, the e-mail is spam.

To explain the Naïve Bayes method, this example is discussed as follows:

Figure 2 Separating non-spam mails from spams

P(Ci) : P(Class=Spam) = S / N

P(Ci) : P(Class=Ham) = H / N

The total occurrence of the word Free in spam

$$X1 = \frac{A}{Spam}$$

$$\sum_{i=0}^{n} Free(Ham) = B$$

$$X2 = \frac{B}{Ham}$$

After obtaining the threshold (X1, X2), if the number of occurrences of the word Free in 21st spam email is closed to the X1, that mail is spam and if it is closed to X2, it is HAM. The same operations on other components can be checked [7].

Using SVM (Linear support vector machine) in classification issues is a new approach that in recent years has become an attractive subject for many and is used in a wide range of applications, including OCR, handwriting recognition, signs recognition and so on. SVM approach is that in the training phase, they try to maintain the decision boundary in such a way that its minimum distance to each of the considered categories becomes maximum distance. This choice makes the decision practically to tolerate the noisy environment and have a good response. This boundary selection method is based on the points called support vectors. Thus in the training phase, general characteristics of spam are extracted according to the data analysis obtained from data collection and training is done based on it. Then testing phase was performed based on the mentioned cases and each time compared with original data until the results became optimized [8]. The proposed method is as follows: first measurement criteria for spam are determined which contains implicit (non-material) and explicit (content). Implicit cases that are analyzed by decision tree include protocol type, content length, content type, time, frequency, receiver number, etc. explicit cases are determined by Naïve Bayes and support vector, such as the repetition of words Victory, Win, three zeros in a row, and so on. In fact, the desired dataset is a combination of implicit properties inside the decision trees and explicit characteristics used in Naïve Bayes method and vector machine and the results of all three methods are surveyed by voting. In other words, the implicit characteristics of the data set were analyzed by decision tree algorithm and the obtained results are completed through fuzzy - Mamdani rules [9]. Then explicit characteristics in Naïve Bayes rule and support vector machine are evaluated and the results of all three methods are surveyed by voting. Each of the mails inside the data collection are entered into Naïve Bayes, support vector machine and decision tree. If at least two of the three proposed algorithms were determined correctly, or in other words, at least two of the three proposed algorithms are like minded, the result is acceptable. In this method, results are divided in two groups: high reliability for all three algorithms being likeminded and average reliability for two of the three proposed algorithms being likeminded. Ultimately in the final test, data set was divided into four parts by K-fold method and the first quarter was analyzed for testing and the rest for learning; next, the second quarter of data set was analyzed for testing and first, third and fourth quarters were analyzed for learning; then the third quarter was analyzed for testing and first, second and fourth quarters were analyzed for learning; also the fourth quarter was considered for testing and first, second and third quarters were considered for learning [10].

4. RESULT AND DISCUSSION

Dataset for implementing the proposed method contain 1000 emails in which 300 (30%) are spam and 700 (70 %) are non-spam. In the last column of this data set, there is a (Class) column and inserting number 1 in this column means spam and inserting zero means non-spam for every mail. Examples of keywords for the explicit section regarding the implementation of the Naïve Bayes method and support vector machine include:

Victory, Money, Win, Lottery, 000, ###,…

Another section of this data set containing the implicit characteristics may be used to implement the decision trees, including:

Sending time, type of text, text length, frequency, receiver number, sender number and...

The purpose of testing the proposed data set is to assess the accuracy of detection of the proposed method and showing better spam detections compared with Naïve Bayes method, support vector machine, or decision tree [11]. After analyzing the data set inside the Naïve Bayes method and extracting

efficiency and accuracy percentages and bright-dark spots, that same data set inside the decision tree and support vector machine is analyzed and its accuracy and efficiency percentages is calculated and the results will be analyzed by voting method. Then results with the like-mindedness of at least two of the three proposed algorithms algorithm are determined as the final results. To demonstrate the efficiency, f the proposed method is performed by one of the methods discussed [12]. F-Measure and accuracy criteria were compared so that the testing data set is divided into ten parts and hundred, two hundreds, three hundreds to thousand mails were tested. The results will be compared with the results of spam swarm optimization method that include negative selection algorithm and particle swarm optimization on measurement and accuracy metrics.

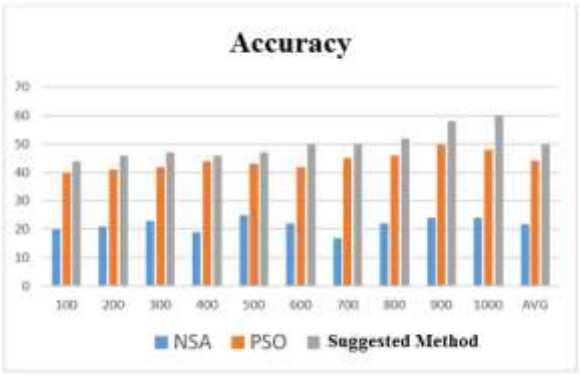


Figure 3 Accuracy Compare Between Methods

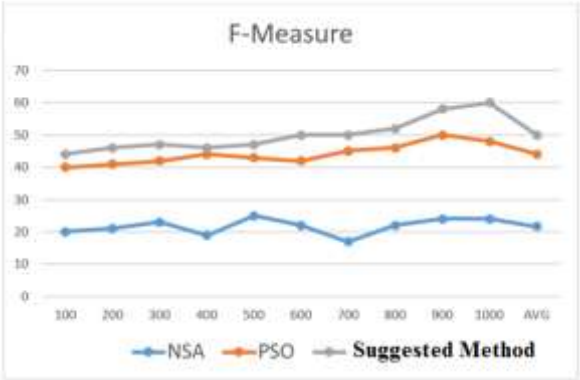


Figure 4 F-Measure

5. CONCLUSION

This method provides a new approach to detect spam using a combination of decision tree, Naïve Bayes method, support vector machine method and surveying by voting to extract the behavioral patterns of spam. Since there are no entirely clear features in the real world, degrees below horizon is normal and reasonable to explain the behavioral characteristics [13]. Decision tree begins identifying the mail and spams in dataset using fuzzy - Mamdani rules. Then Naive Bayes method is performed on the selected data set through Bayes formula of this operation. Then support vector machine algorithm analyzes the explicit data and voting method by the obtained percentages from all three methods and fining like-mindedness of at least two of the three proposed algorithms, shows the accuracy of detection. Proposed method not only shows a better performance compared with the independent use of each of the three methods, but divide's the detection into two groups of almost reliable and very reliable by using the majority of votes in detecting spam and normal mails (common TP and TN of all three methods).

One of the most important things in determining the optimal method to detect spam is to minimize the number of non-spams mails that are known as spam, since finding and deleting spams between safe e-mails known is simple for the user while finding a safe mail between spams is usually difficult and time consuming. To improve the accuracy of spam detection results, three methods were used and better statistics were provided for spam detection through voting method. Comparing the proposed approach with some of previous methods that have already been done show a better performance regarding the accuracy of results. Adding a fuzzy preprocessing level for processing email contents for user was done by using mail classifications into categories based on content, subject, sender, sending time, receiver number, sender number, etc. and the integration of Naïve Bayes method, decision tree and support vector machine based on implicit and explicit components of a mail for all the three methods and classification was done by voting. Using false positive and negative rates increased the accuracy of statistical filters for spam detection accuracy and lowered detection error rate..

6. RECOMMENDATIONS AND FUTURE WORK

To improve the performance of proposed method in the future, more details can be evaluated by the development of decision tree. For example, more detailed non-content cases including: sending time, sending protocol, content length, content type, time zone, receiver number, frequency and number of attachments can increase the accuracy of decision tree in spam detection. For Naïve Bayes and support vector machine methods, adding content details, some parts of a mail such as: subject, content, sender, keywords and user interests, results in improved performance of these two methods regarding the classification of spam and non-spam mails. Finally, using all the three methods in voting algorithm will show a better efficiency percentage. More K-fold divider and learning percentage increases proposed method's accuracy of detection. In other words, K-fold amount and learning percentage considered for the proposed method have a direct relationship with the accuracy of detection. Of course, if details and K-fold amount exceeds a certain extent, implementing the method will be more complex.

7. REFERENCES

[1]. Wu, C.T., Cheng, K.T., Zhu, Q., and Wu, Y.L., 2008, "Using Visual Features For Anti-Spam Filtering", In Proceedings of the IEEE International Conference on Image Processing, Vol. 29, Iss. 1, pp. 63-92.

[2]. Goodman, J., and Rounthwaite, R., 2004, "Stopping Outgoing Spam", In Proceedings of the 5th ACM Conference on Electronic Commerce, pp. 30-39.

[3]. Siponen, M., and Stucke, C., 2006, "Effective Antispam Strategies In Companies: An International Study", In Proceedings of the 39th IEEE Annual Hawaii International Conference on Transaction on Spam Detection, Vol. 6, pp. 245-252.

[4]. Cody, S., Cukier, W., and Nesselroth, E., 2006, "Genres Of Spam: Expectations And Deceptions", In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Vol. 3, pp. 48-51.

[5]. Golbeck, J., and Hendler, J., 2006, "Reputation Network Analysis For Email Filtering", In Proceedings of the First International Conference on Email and Anti-Spam, pp. 21-23.

[6]. Liang, Z., Jianmin, G., and Jian, H., 2012, "The Research and Design of an Anti-open Junk Mail Relay System", In Proceedings of the First IEEE International Conference on Computer Science and Service System, pp. 1258-1262.

[7]. Feamster, N., and Ramachandran, A., 2006, "Understanding The Network-Level Behavior Of Spammers", In Proceeding of the 3th ACM Conference on Email and Anti-Spam, Vol. 36, Iss. 4, pp. 291-302.

[8]. Lili, D., and Yun, W., 2011, "Research And Design Of ID3 Algorithm Rules-Based Anti-Spam Email Filtering", In Proceedings of the Second IEEE International Conference on Software Engineering and Service Science, pp. 572-575.

[9]. Zhitang, L., and Sheng, Z., 2009, "A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree", In Proceedings of the Ninth IEEE International Conference on Computer and Information Technology , Vol. 2, pp. 236-241.

[10]. Duquenoy, P., Moustakas, E., and Ranganathan, E., 2005, "Combating Spam Through Legislation: A Comparative Analysis Of Us And European Approaches", In Proceedings of the Second International Conference on Email and Anti-Spam,pp. 15-22.

[11]. Jones, L., 2007, "Good Times Virus Hoax FAQ", Available: http://cityscope.net/hoax1.html, [Accesed: Jul. 10, 2015].

[12]. Singhal, A., 2007, "An Overview Of Data Warehouse, Olap And Data Mining Technology", Springer Science Business Media, LLC, Vol. 31, pp. 19-23.

[13]. Ismaila, I., and Selamat, A., 2014, "Improved Email Spam Detection Model With Negative Selection Algorithm And Particle Swarm Optimization", Elsevier Journal of Alliance and Faculty of Computing, Vol. 22, pp. 15-27.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 137 - 140, 2016, ISSN:- 2319–8656

Vehicular Messaging In IOT Using Epidemic Routing

Vrushali Pavitrakar
Computer Department PVPIT,
Pune, India

Navnath Kale
Computer Department PVPIT,
Pune, India

Abstract: Now a days there are lots of inventions done in vehicles and some may in progress, these inventions help to resolve the challenges caused by the increasing transportation issues. These modern vehicles equipped with a large amount of sensors, actuators and communication devices e.g. GPS. New vehicles have possessed powerful sensing, networking, communication and data processing capabilities and can communicate with other vehicles or exchange information with the external environments over various protocols. This can be done with the help of Vehicular communication systems; these are the networks in which vehicle and roadside units are the communicating node, with each other provide the information, such as safety warnings and traffic information. They can be effective in avoiding accidents and traffic congestion. To solve these issue sending and receiving messages is an important factor. All this things are good when the number of vehicles are less, system must be able to handle traffic spike or sudden demands caused by special events or situations such as sport games or emergencies. This is nothing but a scalability challenge which is going to overcome in this paper using routing protocol called Epidemic routing protocol.

Keywords: VCS (Vehicle communication system). VANET (Vehicular Ad Hoc Network), GPS (Global Positioning System), IOT (Internet of Things)

1. INTRODUCTION

The inventions in IoT have provided a promising opportunity to address the increasing transportation issues such as heavy traffic, congestion and vehicle safety. (What is IOT? The Internet of Things (IoT) is the network of physical objects, devices, vehicles, buildings and other items which are embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data.) Now a day's vehicle builds up with variety of devices like sensors, actuators and communication devices. Vehicular communication systems are networks in which vehicles and roadside units are the communicating nodes, providing each other with information, such as safety warnings and traffic information. They can be effective in avoiding accidents and traffic congestion. The main incentive for VCS is safety and eliminating the excessive cost of traffic collisions. According to World Health Organizations (WHO), road accidents annually cause approximately 1.2 million deaths worldwide. If preventive measures should be taken otherwise road death can be the leading cause of death. In 2001, it was mentioned in a publication that ad hoc networks can be formed by cars and such networks can help overcome blind spots, avoid accidents, etc. This issue can be solve when vehicle talk each other. It is very shocking when we say vehicle can communicate with each other, but yes now a days it is possible. What is vehicle to vehicle communication? - V2V is a technology designed to allow automobiles to "talk" to each other. V2V is currently in active development by General Motors. Now a day's V2V is possible with BMW, Daimler, Honda, Audi, and Volvo. V2V is also known as VANETs. Sometimes around 2020, cars will communicate with each other and alert drivers to roadside hazards ahead. V2V communication technology help to improve safety by allowing vehicles to "talk" to each other and avoid many crashes altogether by exchanging basic safety data, such as speed and position.

With safety data from nearby vehicles, vehicles can identify risks and provide drivers with warnings to avoid other vehicles in common crash types such as rear-end, lane change, and intersection crashes. In this case IoT has received a lot of attention and is expected to bring benefits to numerous application areas including health care, manufacturing and transportation.

2. LITERATURE SURVEY

All Wireless technology plays a vital role in vehicular networks. The original idea is the roadside infrastructure and the radio-equipped vehicles could communicate using wireless networks. For more effective routing, Vehicular Ad- hoc Networks (VANET) has been developed. VANETs primarily designed to support the communication between different vehicles (V2V) and the communication between vehicles and the roadside infrastructures (V2I) [2]. VANET applications focused on improving driver's safety and offered functions such as traffic monitoring and update, emergency warning and road assistance [3].

Now a day's all modern cars equipped with internet help to bring all these ideas. Olariu, Khalil, and Abuelela [3] propose to integrate vehicular networks, sensors, and on-board devices in vehicles to provide more flexibility. The integration of sensors and communication technologies help us to track the changing status of vehicle. IoT explains a future in which a variety of physical objects and devices around us such sensors, radio frequency identification (RFID) tags, GPS devices, and mobile devices will be associated to the Internet and allow these objects and devices to connect, cooperate and communicate within social, environmental, and help to user to reach common goals [4,5]. Speed & Shingleton [6] propose an idea to use the "unique identifying properties of car registration plates" to connect various things. IoT technologies make it possible to track each vehicle existing location, monitor its movement and predict it future location. An intelligent informatics system (iDrive system) developed by BMW used various sensors and tags to monitor the environment such as tracking the vehicle location and the road condition to provide

driving directions [7]. Leng and Zhao [8] propose an intelligent internet-of-vehicles system (known as IIOVMS) to collect traffic information from the external environments and to monitor and manage road traffic in real time. Here we proposed a routing protocol which will eventually deliver lots of messages. The epidemic routing protocol able to deliver all messages to its destination without any failure. Epidemic Routing [9] is to distribute messages to hosts, called carriers, within connected portions of ad hoc networks.

3. SYSTEM ARCHITECTURE

3.1 Epidemic Routing System

Routing protocols allow nodes with wireless adaptors to communicate with other without any pre-existing network infrastructure. Rapidly changing network topology help to deliver messages in the case where there is never a connected path from source destination. Here we are using Epidemic Routing, where random pair-wise exchanges of messages among mobile hosts ensure eventual message delivery. The goals of Epidemic Routing are to: I) maximize message delivery rate, ii) minimize message latency and iii) minimize the total resources consumed in message delivery. Wireless network adaptors in portable computing devices, such as cellular phones, personal digital assistants, and laptops. The goal of this type of protocol is to develop techniques for delivering application data with high probability even when there is never a fully connected path between source and destination or if there is lots of messages. In this way, messages are quickly distributed through connected portions of the network. Epidemic Routing then relies upon carriers coming into contact with another connected portion of the network through node mobility. At this point, the message spreads to an additional island of nodes. Through such transitive transmission of data, messages have a high probability of eventually reaching their destination.
Figure 1 shows Epidemic Routing at a high level, with mobile nodes represented as dark circles and their wireless communication range shown as a dotted circle extending from the source. In Figure 1(a), a source, S, wishes to send a message to a destination, D, but no connected path is available from S to D. S transmits its messages to its two neighbors, C1 and C2, within direct communication range. At some later time, as shown in Figure 1(b), C2 comes into direct communication range with another host, C3, and transmits the message to it. C3 is in direct range of D and finally sends the message to its destination.

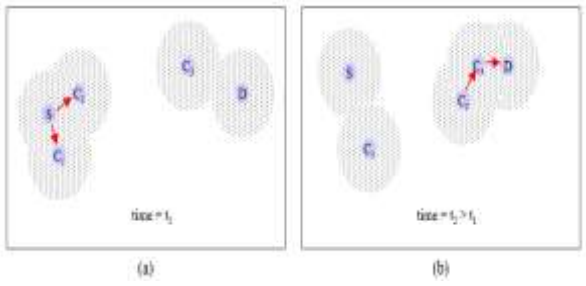


Fig1. A source, S, wishes to transmit a message to a destination but no connected path is available in part (a). Carriers, C1-C3 are leveraged to transitively deliver the message to its destination at some later point in time as shown in (b).

The goal of Epidemic Routing is: to deliver a message (update) with high probability to a particular host. In fact if there is certain situation where message count is high than regular, in this case also routing protocol should be able to send or receive message eventually. Likewise Epidemic Routing should be able to message broadcast/multicast in partially connected ad hoc networks. The overall goal of Epidemic Routing is to maximize message delivery rate and minimize message delivery latency, while also minimizing the aggregate system resources consumed in message delivery. We explore message delivery rate and resource consumption under a number of different scenarios. Our results show that Epidemic Routing is able to deliver all messages in where existing ad hoc routing protocols fail to delivery some messages because of limited node connectivity. Epidemic Routing delivers 100% of messages assuming enough per-node buffering to store between 10-25% of the messages originated in the scenario.

1. Goal:- The goals of Epidemic Routing are to:
   - Efficiently distribute messages through partially connected ad hoc networks in a probabilistic fashion,
   - Minimize the amount of resources consumed in delivering any single message
   - Maximize the percentage of messages that are eventually delivered to their destination

2. Epidemic Routing Protocol: Epidemic Routing supports the eventual delivery of messages based on minimal assumptions like only periodic pair-wise connectivity is required to ensure eventual The Epidemic Routing protocol works as follows.

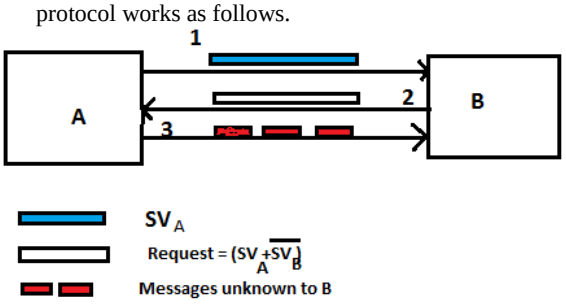


Figure 2: The Epidemic Routing protocol when two hosts, A and B, come into transmission range of one another

Each host maintains a buffer consisting of messages that it has originated as well as messages that it is buffering on behalf of other hosts. For efficiency, a hash table indexes this list of messages, keyed by a unique identifier associated with each message. Each host stores a bit vector called the summary vector that indicates which entries in their local hash tables are set. When two hosts come into communication range of one another, the host with the smaller identifier initiates an anti-entropy session. To avoid redundant connections, each host maintains a cache of hosts that it has spoken with recently. Anti-entropy is not re-initiated with remote hosts that have been contacted within a time period. During anti-entropy, the two hosts exchange their summary vectors to determine which messages stored remotely have not been seen by the local host. In turn, each host then requests copies of messages that it has not yet seen. The receiving host maintains total autonomy in deciding whether it will accept a message. For example, it may determine that it is unwilling to carry messages larger than a given size or destined for certain hosts. We do model a maximum queue size associated with each host, which determines the maximum number of messages a host is willing

to carry on behalf of other hosts. Figure 2 depicts the message exchange in the Epidemic Routing protocol. Host A comes into contact with Host B and initiates an anti-entropy session. In step one, A transmits it summary vector, and SVA to B. SVA is a compact representation of all the messages being buffered at A. Next, B performs a logical AND operation between the negation of its summary vector, SVB, and SVA. That is, B determines the set difference between the messages buffered at A and the messages buffered locally at B. It then transmits a vector requesting these messages from A. In step three, A transmits the requested messages to B. This process is repeated transitively when B comes into contact with a new neighbor.

3.2 Proposed System Architecture

Given sufficient buffer space and time, these anti-entropy sessions guarantee eventual message delivery through such pair-wise message exchange. Epidemic Routing associates a unique message identifier, a hop count, and an optional acknowledgment request with each message. Thus, high priority messages might be marked with a high hop count, while Given that messages are delivered probabilistically in epidemic routing, certain applications may require acknowledgments of message delivery.

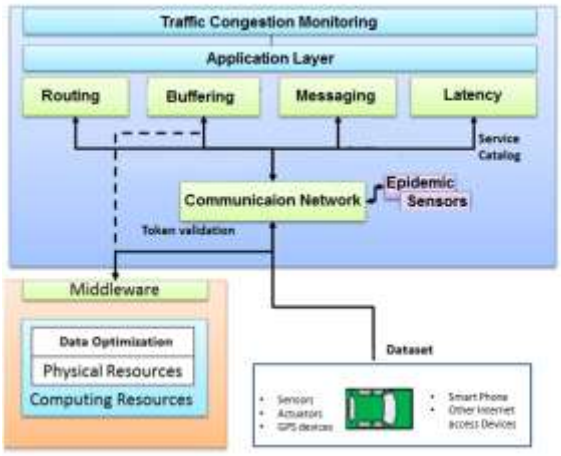


Fig 3: System Architecture

Each host sets a maximum buffer size that it is willing to allocate for epidemic message distribution. The buffer size limits the amount of memory and network resources consumed through Epidemic Routing. In general, hosts will drop older messages in favor of newer ones upon reaching their buffer's capacity. To ensure eventual delivery of all messages, the buffer size on at least a subset of nodes must be roughly equal to the expected number of messages in transit at any given time. Otherwise, it is possible for older messages to be deleted from all buffers before delivery. Fig 3 shows system architecture diagrammatically.

4. CONCLUSION

In this paper, we are going to use epidemic routing protocol which will allow confirm message delivery even if there are 100 or 1000 of messages to be deliver in some special cases. Existing ad hoc routing protocols are robust to rapidly changing network topology, but they are unable to deliver packets in the presence of a network partition between source and destination.

Including mobile sensor networks and disaster recovery scenarios, nodes can be spread over wide geographical distances such wide dispersion makes it unlikely that a connected path can always be discovered, because of that it is impossible to perform message delivery using current ad hoc routing protocols. Thus, we introduce Epidemic Routing, where random pair-wise exchanges of messages among mobile hosts ensure eventual message delivery. The goals of Epidemic Routing are to maximize message delivery rate and to minimize message latency while also minimizing the total resources (e.g., memory and network bandwidth) consumed in message delivery. In our case we will show that Epidemic Routing delivers 100% of eventual message delivery with reasonable resource consumption where existing ad hoc routing protocols are unable to deliver some messages because no end-to-end routes are available.

ACKNOWLEDGMENTS

I wish to thanks my guide Mr. Navnath kale sir for their guidance and encouragement for this work. I am also thankful to the principle of PVPIT College for providing me the opportunity to embark this project.

5. REFERENCES

[1] Wu He, Gongjun Yan and Li Da Xu, Senior Member, IEEE, "Developing Vehicular Data **Cloud** Services in the IoT Environment.

[2] Wu He, Gongjun Yan and Li Da Xu, Senior Member, IEEE, "Developing Vehicular Data **Cloud** Services in the IoT Environment.

P. Papadimitratos, A. La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation," IEEE Communications Magazine, 47(11), 84-95, 2009.

[3] S. Olariu, I. Khalil, and M. Abuelela, "Taking VANET to the clouds," International Journal of Pervasive Computing and Communications, 7(1), pp.7-21, 2011s

[4] European Commission Information Society. Internet of Things in 2020: a Roadmap for the Future. Available from: www.iot-visitthefuture.eu

[5] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Computer Networks, 54(15), pp. 2787-2805, 2010.

[6]]C. Speed, and D. Shingleton, "An internet of cars: connecting the flow of things to people, artefacts, environments and businesses," In Proceedings of the 6th ACM workshop on Next generation mobile computing for dynamic personalised travel planning, pp. 11-12, ACM, 2012.

[7] E. Qin, Y. Long, C. Zhang, and L. Huang, "Cloud Computing and the Internet of Things: Technology Innovation in Automobile Service," In Human Interface and the Management of Information.

Information and Interaction for Health, Safety, Mobility and Complex Environments (pp. 173-180). Springer Berlin Heidelberg, 2013.
[8] Y. Leng, and L. Zhao, "Novel design of intelligent internet-of-vehicles management system based on cloud-computing and Internet-of-Things," Proceedings of 2011 International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT), Vol. 6, pp. 3190-3193, 2011
[9] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic Algorithms for Replicated Database Maintenance. In Proceedings of the Sixth Symposium on Principles of Distributed Computing, pages 1–12, August 1987

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 141 - 146, 2016, ISSN:- 2319–8656

Holistic Approach for Arabic Word Recognition

Talaat M. Wahbi
College of Computer Science and Information Technology
Sudan University of Science and Technology
Khartoum, Sudan

Mohamed E. M. Musa
College of Computer Science and Information Technology
Sudan University of Science and Technology
Khartoum, Sudan

Abstract. Optical Character Recognition (OCR) is one of the important branches. One segmenting words into character is one of the most challenging steps on OCR. As the results of advances in machine speeds and memory sizes as well as the availability of large training dataset, researchers currently study Holistic Approach "recognition of a word without segmentation". This paper describes a method to recognize off-line handwritten Arabic names. The classification approach is based on Hidden Markov models.. For each Arabic word many HMM models with different number of states have been trained. The experiments result are encouraging, it also show that best number of state for each word need careful selection and considerations.
Keywords: pattern recognition; HMM; Holistic approach; offline recognition: Arabic word recognition

1. INTRODUCTION

One of the important branches of pattern recognition is Optical character recognition (OCR). OCR concerns depend on the stages of pattern recognition system. Firstly, it addresses text types like: digits, letters, and words. The second stage is segmentation to letters or strokes in case of using word or continues to the next stage without segmentation (holistic approach). In this approach a word is treated and identified as entity. The third stage is preprocessing which detect the main problems in scanning or writing like skew or noise. Feature extraction stage in OCR is a main issue to classify recognition to online in case of using pen moving direction, pen press,….etc beside the image, or offline in case of image features. The final stage is to use classifier to evaluate the previous stages.

All these stages make different challenges to detect the best parameters for each stage, and this increase when we use a cursive handwritten. For example Arabic words have many letter's shapes, dots, in addition to letter's overlapping. All these difficulties in Arabic language itself let Arabic being late in progress that has been achieved in the field of handwritten word recognition. Another challenge is the lack of special task handwritten datasets.

Arabic names used today have so much repetition, such as names of the prophets (Muhammad, Ibrahim... etc.) and names of the Caliphs and compound names whose first element is Abd (slave of God) (Abdullah, Abdul Rahman... etc.), and there are many examples of repetitive names (such as Adil, Awad ... etc.), together with a few common names. Therefore the idea of designing a system that uses the Holistic Approach to quickly recognize the common names and resort to the use of the Analytical Approach to recognize the names that are not common (Figure 1), is worthy of consideration for the probability of designing an effective system to recognize the names. This paper examines the effectiveness of the first part of this system which is the use of probabilistic neural networks in the inclusive Recognition of the most common Arabic names.

The rest of the paper is organized as follows: Section 2 sketches some related studies in HWR using HMMs. Section 3 briefly introduces HMMs. Section 4 describes in general SUST names dataset. Section 5 illustrates and outlines the results achieved by the experiments performed. Finally, a conclusion is drawn with future work outlooks in section 6.

2. RELATED STUDIES

The application of HMMs to Arabic OCR was first attempted by Amin and Mari[1]. Subsequently Khorsheed and Clocksin [2] present a technique for the offline recognition of cursive Arabic script based on an HMM. AlKhateeb et al. design a word-based off-line recognition system using Hidden Markov Models (HMMs). They extract several structural features and a group of intensity features using a sliding window. Experiments were carried out using the IFN/ENIT database which contains 32,492 handwritten Arabic words [3].Volker Märgner eta al presents the IfN's Offline Handwritten Arabic Word Recognition System. The system uses Hidden Markov Models (HMM) for word recognition, and is based on character recognition without explicit segmentation [4]. Somaya Alma'adeed et al. present a complete scheme for unconstrained Arabic handwritten word recognition based on a multiple hidden Markov models (HMM) [5]. Ramy Al-Hajj and Chafic Mokbel present results of a language independent handwritten recognition baseline system developed to recognize cursive handwritten words. The system is based on a stochastic Hidden Markov Model [6].

3. HIDDEN MARKOV MODEL (HMM)

A hidden Markov model is a stochastic finite state machine, specified by a tuple (S;A;π) where

S is a discrete set of hidden states with cardinality N,

π is the probability distribution for the initial state

$$\pi(i) = P(s_i) \qquad s_i \in \mathcal{S}$$

A is the state transition matrix with probabilities:

$$a_{ij} = P(s_j \mid s_i) \qquad s_i, s_j \in \mathcal{S}$$

Where the state transition coefficients satisfy

$$\sum_{s_j \in \mathcal{S}} a_{ij} = 1\,,\ s_i \in \mathcal{S}$$

The states themselves are not observable. The information accessible consists of symbols from the alphabet of observations O = (o1, ... ,oT) where T is the number of samples in the observed sequence. For every state an output distribution is given as

$$b_i(k) = P(o_t = k \mid s_i) \qquad k \in \mathcal{O},\ s_i \in \mathcal{S}$$

Thus, the set of HMM parameters θ consists of the initial state distribution, the state transition probabilities and the output probabilities. MMs can be used for classification and pattern recognition by solving the following problems:

The Evaluation Problem: Given the model with parameters θ, calculate the probability for an observation sequence O, Let O=(o1,… ,oT) denote the observation sequence and S=(s1,… ,sT) ; a state sequence. The probability P(O|θ) can be obtained by Forward Algorithm.

The Decoding Problem: Find the optimal state sequence for an observation sequence argmaxSϵST P(S|O,θ). This can be done by the Viterbi algorithm [7].

The Learning Problem: Given an observation sequence O and the HMM parameters, find the parameters Ө which maximize P(O|θ) i.e. Ө argmaxθ P(O|θ). This question corresponds to training an HMM. The state sequence is not observable. Therefore, the problem can be viewed as a missing-data problem, which can be solved via an EM-type algorithm. In the case of HMM training [7], this is the Baum-Welch algorithm. A tutorial on HMM models, the estimation problems mentioned above, and their applications to modeling a recognition system can be found in [7].

4. SUST NAMES DATASET

Arabic males names data set was used to detect the efficiency of the suggested comparison, it's a new dataset publish by SUST ALT group, it contain about 40,000 sample for 40 common males and females name in Sudan (this statistical depend on a previous dataset from the same group), figure (1) below shows the form used in the data collection process.

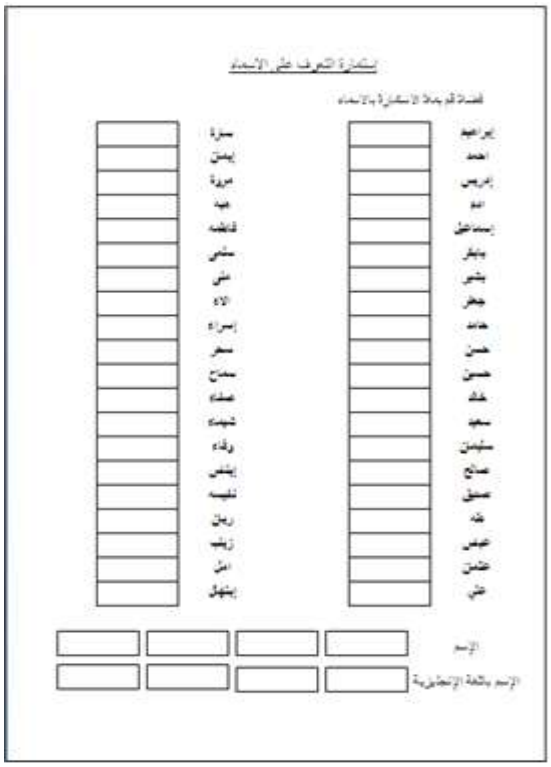


Figure 1: SUST names dataset

5. EXPERIMENT AND RESULTS

The main Recognition system stages are: Preprocessing, framing, features extraction, vector quantization, classification. We choose males names from SUST dataset with 100 samples per class for training and 50 samples per class for testing, figure (2) show all processes in details as follows.

5.1. Preprocessing

This stage contain many sub stages:

- Noises remove.
- Cropping and binarization: extract just handwritten word image and get the binary image.
- Resizing: to cope image dimensions difference in size, forming all images size to be 60X140.

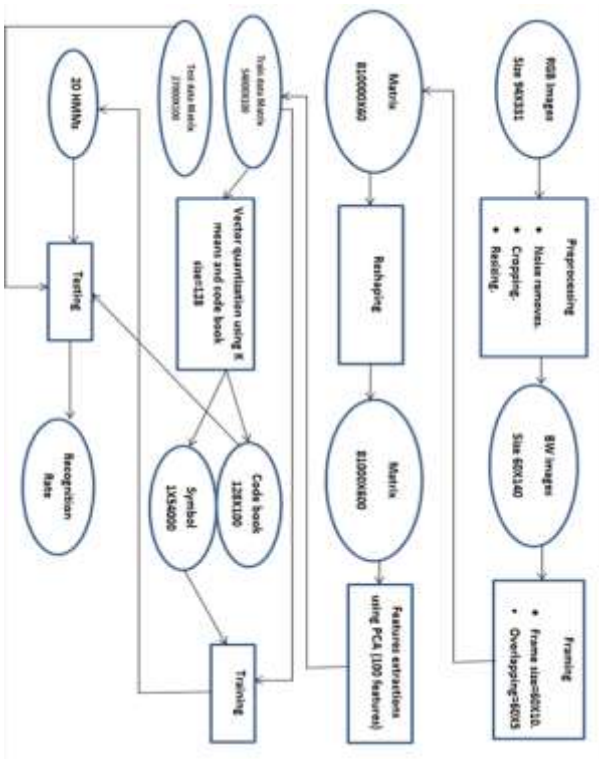


Figure (2): show general view of recognition system

5.2. Framing

Any vector split into frames by window size equal to half frame and frame size equal to 60X10 pixels. The choosing of this size depends on experiments.

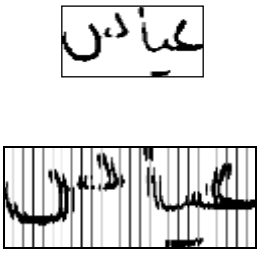


Figure (3): Upper image show the image after resized and the lower show image after framing.

5.3 Reshaping

Each frame reshape in to vector 1X600

5.4 Features extraction

Three testing done to choose the best features number using PCA, 100 features founded to been the best one.

5.5 Vector quantization

Codebook generated using k means clustering algorithm with 128 clusters, this number is the dominated in many researches [8, 9].

5.6 Classification

Model Discriminant HMM (MD-HMM) is used. The main goal of classification is to address states number effect in the HMM recognition system. 20 HMMs trained using Viterbi, states number set is {3,4,5,6,7} and their recognition rates shown in table (1).

Table (1): train and test recognition rates for states set

states	train	test
3	78.2%	52.1%
4	84.2%	54.3%
5	88.3%	56.6%
6	91.35%	59%
7	92.4%	63%

6. TEST CONFUSION MATRICES

6.1 In case 3 states

Table (2): three states model confusion matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	27	1	3	0	3	0	0	1	2	0	4	0	0	1	2	1	1	0	1	3
2	2	23	1	3	8	0	2	0	1	0	1	0	0	0	2	3	4	0	0	0
3	5	0	26	1	0	9	1	0	1	0	1	0	0	2	1	1	0	0	2	0
4	1	3	2	17	10	5	1	2	1	1	1	0	0	1	3	0	1	0	0	1
5	0	3	0	3	19	0	0	7	1	3	2	2	0	0	1	0	9	0	0	0
6	3	0	20	0	0	19	3	0	2	0	0	0	0	0	0	0	2	0	0	1
7	1	0	6	2	1	5	23	0	0	0	8	0	0	3	0	0	1	0	0	0
8	2	0	1	0	4	0	0	31	4	1	3	0	0	0	1	0	2	0	1	0
9	1	1	1	0	0	0	0	0	41	0	0	1	0	1	1	0	0	2	0	1
10	2	3	1	0	7	0	2	0	0	22	4	0	1	0	1	0	1	3	0	3
11	8	0	1	0	2	0	3	6	0	0	29	0	0	0	0	0	0	0	0	1
12	0	0	0	1	1	0	0	1	1	2	2	29	0	5	2	2	0	2	1	1
13	2	0	5	0	1	3	2	1	2	0	0	0	32	1	1	0	0	0	0	0
14	5	0	4	1	0	0	4	1	0	0	0	1	0	25	0	6	0	1	2	0
15	6	0	1	2	5	1	1	3	1	0	2	0	1	1	22	1	2	1	0	0
16	2	2	0	1	0	2	4	0	0	0	3	0	0	2	0	27	5	0	1	1
17	0	4	0	3	6	0	0	0	1	1	4	0	0	0	0	4	26	0	1	0
18	2	0	2	0	1	0	0	0	0	3	0	0	0	2	0	3	0	27	6	4
19	4	0	1	0	0	0	0	1	0	0	3	0	0	0	0	3	0	9	28	1
20	6	0	0	0	0	1	0	2	1	0	6	1	0	1	0	2	0	1	1	28

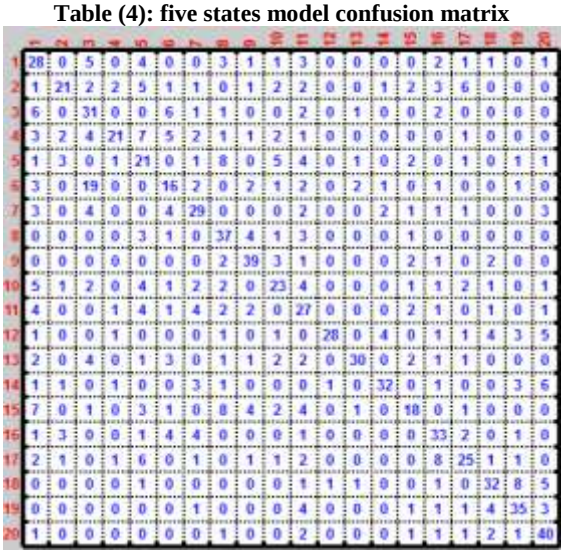
6.2 In case 4 states

Table (3): four states model confusion matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	24	0	3	0	4	0	0	0	0	0	2	0	1	0	4	3	2	1	2	4
2	2	27	2	2	6	0	1	0	1	1	4	0	0	0	1	2	1	0	0	0
3	6	0	26	0	0	6	2	0	3	0	1	0	2	2	1	0	0	1	0	0
4	2	2	5	19	6	4	4	2	1	1	1	0	0	0	2	0	1	0	0	0
5	2	2	0	0	19	0	0	8	0	3	5	0	1	0	2	0	6	0	1	1
6	3	0	13	0	0	20	7	0	1	3	1	0	0	0	1	0	1	0	0	0
7	3	0	5	1	0	3	30	0	0	0	4	0	0	1	0	0	0	0	0	3
8	0	0	0	1	2	0	1	35	0	3	1	0	0	0	4	1	0	2	0	0
9	1	0	0	0	0	0	1	5	36	1	1	1	0	0	1	0	1	1	0	1
10	4	1	2	0	5	0	1	1	2	26	3	0	0	0	2	0	1	1	0	1
11	10	0	1	0	0	1	2	6	3	1	22	0	0	0	1	0	1	0	0	2
12	0	0	0	0	0	0	0	3	1	1	1	29	1	3	0	1	0	5	2	3
13	1	0	3	0	1	3	0	0	1	2	1	0	35	0	2	0	0	1	0	0
14	2	1	1	1	0	0	4	1	0	0	0	0	1	29	0	4	0	0	4	2
15	6	0	1	1	3	1	2	7	3	1	2	0	0	0	19	0	2	1	1	0
16	3	1	3	0	0	0	1	0	0	0	1	0	0	2	2	32	4	0	0	1
17	3	1	0	2	10	0	1	0	0	1	1	0	0	0	2	5	23	0	1	0
18	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	4	0	30	6	6
19	1	0	1	0	0	0	1	3	1	1	3	1	0	0	3	3	0	2	27	3
20	2	0	1	0	0	0	0	2	0	1	4	0	0	0	2	3	0	0	0	35

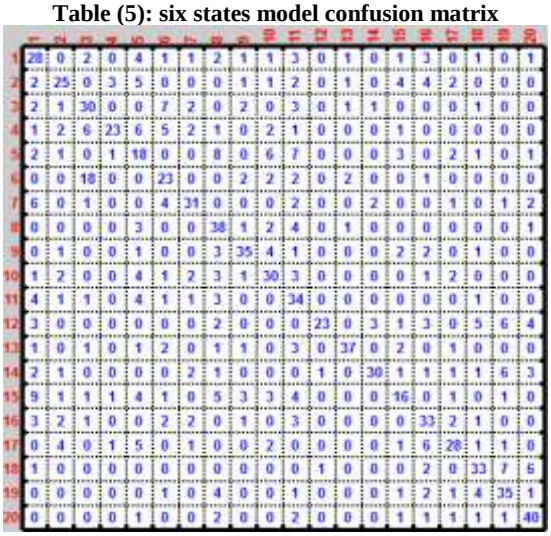
6.3 In case 5 states

Table (4): five states model confusion matrix



6.4 In case 6 states

Table (5): six states model confusion matrix



6.5 In case 7 states

Table (6): seven states model confusion matrix

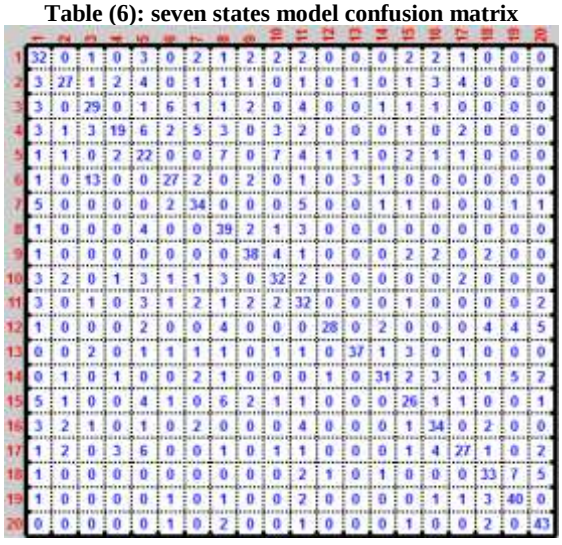


Table (7): explain names and theirs letters numbers

Name/class		Letters no.
1. Abass	عباس	4
2. Hesn	حسن	3
3. Khalid	خالد	4
4. Osman	عثمان	5
5. Saiad	سعيد	4
6. Salih	صالح	4
7. Sedig	صديق	4
8. Soliman	سليمان	6
9. Taha	طه	2
10. Adm	آدم	3
11. Ahmed	أحمد	4
12. Ali	علي	3
13. Babekir	بابكر	5
14. Bsheer	بشير	4
15. Ebraheem	إبراهيم	7
16. Edrees	إدريس	5
17. Esmaeail	إسماعيل	7
18. Gafr	جعفر	4
19. Hamed	حامد	4
20. Hesain	حسين	4

The following issues are observed from the above confusion matrices:

- **States increasing and number of letters**

The Table(8) displays the best number of states for different classes.

Table (8): explain classes and state/states according to best recognition rates

Name/class	States	Name/class	states
1. Abass	7	11. Ahmed	6
2. Hesn	4,7	12. Ali	3,4
3. Khalid	5	13. Babekir	6,7
4. Osman	6	14. Bsheer	5
5. Saiad	7	15. Ebraheem	7
6. Salih	7	16. Edrees	7
7. Sedig	7	17. Esmaeail	6
8. Soliman	7	18. Gafr	6,7
9. Taha	3	19. Hamed	7
10. Adm	7	20. Hesain	7

The above table display clearly the effect of numbers of letters in recognition, especially in short names like Ali and Taha (short names has best recognition rates with small number of states). The rest of names (medium and high) have higher recognition rates with higher number of states.

From Table (9) we note that five names which have highest confusion with other names are (greater than 30 samples): Abass, Khalid, Saiad, Soliman, and Ahmed. All these names have 4 letters except (Soliman). Also these name have high recognition rates with one class except (Saiad).

- **The relation between states and error rates**

Table (10) shows the names which six or more confused samples

The major confused classes are: Saiad with Soliman, salih with Khalid, and Gafr with Hamed and not vice versa. Overlapping play the main role in this confusion as shown in Figure (5).

Table (9): The relation between increasing the number of states and numbers of samples confused with others classes

Name/ Class	3 states model	4 states model	5 states model	6 states model	7 states model
1. Abass	52	52	41	37	36
2. Hesn	17	8	11	16	10
3. Khalid	49	42	41	31	22
4. Osman	17	8	7	6	9
5. Saiad	49	37	39	38	38
6. Salih	26	18	27	25	16
7. Sedig	23	28	22	13	19
8. Soliman	25	38	31	35	33
9. Taha	18	17	17	13	13
10. Adm	11	20	21	32	22
11. Ahmed	44	37	40	41	37
12. Ali	5	3	2	2	3
13. Babekir	2	6	6	6	5
14. Bsheer	20	8	8	6	7
15. Ebraheem	15	30	15	18	19
16. Edrees	26	26	26	26	18
17. Esmaeail	28	20	19	14	13
18. Gafr	19	15	16	18	15
19. Hamed	16	17	19	23	17
20. Hesain	17	27	26	19	18

Table (10): explain classes and state/states according to best recognition rates

Name/ class	3 states model	4 states model	5 states model	6 states model	7 states model
1. Abass					
2. Hesn	Saiad				
3. Khalid	Salih			Salih	
4. Osman	Khalid		Saiad		
5. Saiad	Soliman, Esmaeail	Soliman	Soliman	Soliman, Ahmed	Soliman, Adm
6. Salih	Khalid	Khalid, Sedig	Khalid	Khalid	Khalid
7. Sedig	Ahmed				
8. Soliman					
9. Taha					
10. Adm	Saiad				
11. Ahmed	Abass	Abass			
12. Ali					
13. Babekir					
14. Bsheer					
15. Ebraheem		Soliman	Gafr	Abass	
16. Edrees					
17. Esmaeail		Saiad	Edrees		
18. Gafr			Hamed	Hamed	Hamed
19. Hamed	Gafr				
20. Hesain					

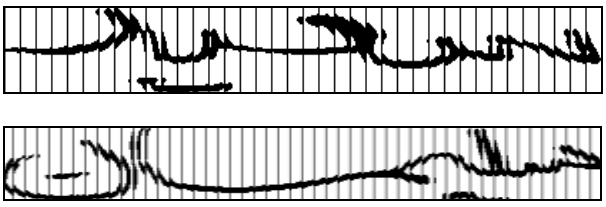


Figure (4): Show overlapping images for (Saiad), and (Soliman).

7. CONCLUSION AND FUTURE WORK

This paper discusses the effects of number of states in a HMM for handwritten Arabic names recognition. According to the results many improvements may achieved. For instance, different set of features may give better results such as Chain code and wavelet code and wavelet. Also adding a post processing component may boost the recognition rate. Another important issue to put in consideration is using multiple states for different words.

8. REFERENCES

[1] Amin and J. Mari, "Machine recognition and correction of printed Arabic text," IEEE Trans. on Systems, Man, and Cybernetics, vol. 19, no. 5, 1989, pp.1300-1306.

[2] M. Khorsheed and W. Clocksin, "Structural Features Of Cursive Arabic Script", The 10th British Machine Vision Conference, University of Nottingham, Nottingham-UK, September-1999.

[3] J.AlKhateeba,J.Rend, J.Jiangb, and H.Al-Muhtaseb Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking, Pattern Recognition Letters,Volume 32 Issue 8, June, 2011 .

[4] V. Maegner, H.El Abed, M.Pechwitz,"Offline Handwritten ArabicWord Recognition Using HMM -a Character Based Approach without Explicit Segmentation" .

[5] S. Almaadeed, C. Higgens, and D. Elliman, "A New Preprocessing System for the Recognition of Off-lineHandwritten Arabic Words", IEEE International Symposium on Signal Processing and Information Technology, December, 2001.

[6] R.Al-Hajj, C.Mokbel," HMM-Based Arabic handwritten cursive reconition system".

[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," Proc. of the IEEE, vol. 77, n. 2, pp 257-285, Feb. 1989.

[8] K.C Jung,S.M Yoon,H.J Kim**,** Continuous HMM applied to quantization of on-line Korean character paces,Pattern Recognition Letters, Volume 21, Issue 4, April 2000, Pages 303-310.

[9] Kenichi Maruyama, Makoto Kobayashi, Yasuaki Nakano, HirobumiYamada. Cursive Handwritten Word Recognition

International Journal of Computer Applications Technology and Research
Volume 5–Issue 3, 147-150, 2016, ISSN:2319–8656

Digital Watermarking Applications and Techniques: A Brief Review

Aaqib Rashid
MCA (Kashmir University)
M.Phil Computer Science (Dr. C.V Raman University)

Abstract: The frequent availability of digital data such as audio, images and videos became possible to the public through the expansion of the internet. Digital watermarking technology is being adopted to ensure and facilitate data authentication, security and copyright protection of digital media. It is considered as the most important technology in today's world, to prevent illegal copying of data. Digital watermarking can be applied to audio, video, text or images. This paper includes the detail study of watermarking definition and various watermarking applications and techniques used to enhance data security.

Index Terms: Watermarking, Techniques, Security, Technology,

I. INTRODUCTION

The advancement of the Internet has resulted in many new opportunities for the creation and delivery of content in digital form. Applications include electronic advertising, real-time video and audio delivery, digital repositories and libraries, and Web publishing. But the important question that arises in these applications is the data security. It has been observed that current copyright laws are not sufficient for dealing with digital data. Hence the protection and enforcement of intellectual property rights for digital media has become a crucial issue. This has led to an interest towards developing new copy deterrence and protection mechanisms. One such effort that has been attracting increasing interest is based on digital watermarking techniques. As steganography pay most attention towards the degree of invisibility, watermarking pay most of its attributes to the robustness of the message and its ability to withstand attacks of removal, such as image operations (rotation, cropping, filtering) etc in case of images being watermarked. Digital watermarking is the process of embedding information into digital multimedia content such that the information (which we call the watermark) can later be extracted or detected for a variety of purposes including copy prevention and control. Digital watermarking has become an active and important area of research, and development and commercialization of watermarking techniques is being deemed essential to help address some of the challenges faced by the rapid proliferation of digital content.

II. DIGITAL WATERMARKING TECHNOLOGY

As we know the main purpose of both cryptography and steganography is to provide secret communication. However, they are not same. Cryptography hides the content of a secret message from malicious people, where as steganography even conceal the existence of the message. But a new emerging technology known as digital watermarking involves the ideas and theories of different subject coverage, such as signal processing, cryptography, probability theory and stochastic theory, network technology, algorithm design, and other techniques [1]. Digital watermarking hides the copyright information into the digital data through certain algorithm. The secret information to be embedded can be some text, author's serial number, company logo, images with some special importance. This secret information is embedded to the digital data (images, audio, and video) to ensure the security, data authentication, identification of owner and copyright protection. The watermark can be hidden in the digital data either visibly or invisibly. For a strong watermark embedding, a good watermarking technique is needed to be applied. Watermark can be embedded either in spatial or frequency domain. Both the domains are different and have their own pros and cons and are used in different scenario. Fig 1. Shows Digital Watermark embedding process and Fig. 2. Shows watermark detection process.

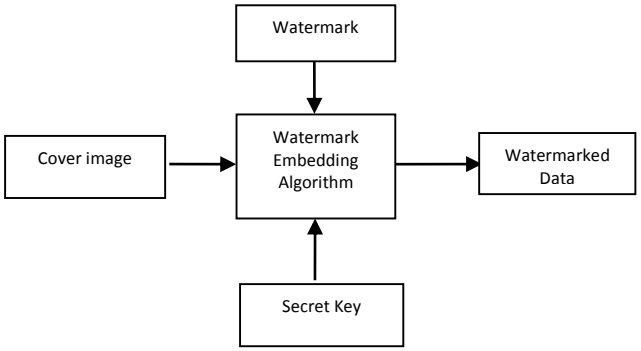


Fig 1. Watermark Embedding Process

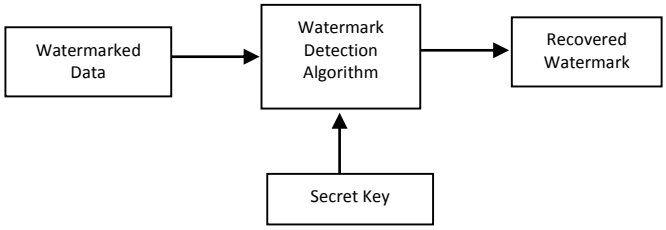


Fig 2. Watermark Detection Process

III. APPLICATIONS

Digital Watermarks are potentially useful in many applications, including:

A. Broadcast Monitoring

Advertisers want to ensure that they receive all of the air time which they purchase from broadcasters (Japan 1997) [5, 40]. A non-technical method in which human observation is used to watch the broadcast and check the originality by seeing or hearing is an error

prone and costly. Thus, there should be an auto-identification system, which may store the identification codes to the broadcast. There are several techniques like cryptography that store the identification code in the file header but the data is unlikely to survive any sort of modifications even format change. Watermarking is obviously a suitable technique for information monitoring. The Watermark exists within the content and is compatible with the installed base of broadcast equipment. Although, embedding the identification code is very complicated compared to the cryptography where the code is placed in the file header. Moreover, it also, affects the visual quality of the work. Still, many companies protect their broadcasts through watermarking techniques.

B. Ownership Assertion

A rightful owner can retrieve the watermark from digital content to prove his ownership. There are limitations with textual copyright notices, as they are easily removable. Copyright notice printed on the physical document cannot be copied along with the digital content. Although, it is possible that text copyright can be placed in an unimportant place of the document to make them unobtrusive [41]. Imperceptible and inseparable watermark is the best solution as compared to the text mark for owner identification. The watermark is not only used for identification of the copyright ownership but for proving the ownership of the document as well. The ownership can be carried out by extracting the embedded information from the watermarked document [42].

C. Transaction Tracking

Transaction tracking is often called fingerprinting, where each copy of the work is uniquely identified, similar to the fingerprint that identifies an individual. The watermark might record the recipient for each legal distribution of the work. The

owner embeds different watermarks in each copy. If the work is misused then will the owner be able to find the traitor? Visible watermarking is adopted for transaction tracking but invisible watermarking is much better. For example, in movie making, the daily videos (also called dailies) are distributed to the persons who are concerned with the movie. Sometimes, the videos are disclosed to the press, so the studios use visible text on corner of the screen, which identifies the copy of dailies. Thus, the watermark is preferred as the text can easily be removed.

D. Content Authentication

The procedure to confirm the integrity of watermarked data and to make sure that the data is not being tampered with i.e. act of establishing or confirming whether image is authentic or not. The term authentication has an extensive range of meanings. For instance, an authority that decides whether a portion of art is authentic or not, can a user view or download it? Finally, the decision is to whether the content of an object is staying intact or not after its transmission on the internet. Many cultural organizations spend time and investing money on new technologies of image documentation and digital libraries construction etc. At the same time, these organizations can guarantee the authenticity of the pieces of art they possess, since they have both the ownership and the experts opinions. When these works of art are digitized and published on the internet, numerous problems take place. Usually several digital images found on the internet have many differences, but at the same time pretending to represent the same piece of art. Use of watermarking related to authentication comprises of trusted cameras, video surveillance and remote sensing applications, digital insurance claim evidence, journalistic photography, and digital rights management systems. Commercially, its applications are expected to grow as does the applications of digital content, for example, GeoVision's GV-Series digital video recorders for digital video surveillance to prevent tampering. The digital work can easily be tampered by using computer resources. A solution to the tamper detection is watermarking, where the authentication mark (watermark) cannot stay with the work after slightest modification. Conversely, the system

does not matter that the work is compressed or significant changes are made. This leads toward semi-fragile watermarking where the system survive the friendly manipulations and fragile against substantial manipulations [5].

E. Copy Control and Fingerprinting

Copy control and fingerprinting are used to prevent people from making illegal copies of the content. This issue is very similar to the transaction tracking of the content. An owner can embed a watermark into digital content that identifies the buyer of the copy (i.e. serial number). If unauthorized copies are found later, the owner can trace the origin of the illegal copies.

IV. DIGITAL IMAGE WATERMARKING TECHNIQUES

In the field of digital watermarking, digital image watermarking has attracted a lot of awareness in the research community for two reasons: one is its easy availability and the other is it convey enough redundant information that could be used to embed watermarks [2]. Digital watermarking contains various techniques for protecting the digital content. The entire digital image watermarking techniques always works in two domains either spatial domain or transform domain. The spatial domain techniques works directly on pixels. It embeds the watermark by modifying the pixels value. Most commonly used spatial domain techniques are LSB. Transform domain techniques embed the watermark by modifying the transform domain coefficients. Most commonly used transform domain techniques is DCT, DWT and DFT. For achieving the robustness and imperceptibility, the transform domain techniques are more effective as compare to the spatial domain.

A. Spatial domain watermarking techniques*:*

The spatial domain represents the image in the form of pixels. The spatial domain watermarking embeds the watermark by modifying the intensity and the colour value of some selected pixels [3]. The strength of the spatial domain watermarking is:

- Simplicity.
- Very low computational complexity.
- Less time consuming.

The spatial domain watermarking is easier and its computing speed is high than transform domain but it is less robust against attacks. The spatial domain techniques can be easily applied to any image. The most important method of spatial domain is LSB.

i. Least Significant Bit (LSB):

The LSB is the simplest spatial domain watermarking technique to embed a watermark in the least significant bits of some randomly selected pixels of the cover image.

The steps used to embed the watermark in the original image by using the LSB [4]:

1) Convert RGB image to grey scale image.
2) Make double precision for image.
3) Shift most significant bits to low significant bits of watermark image.
4) Make least significant bits of host image zero.
5) Add shifted version (step 3) of watermarked image to modified (step 4) host image.

The main advantage of this method is that it is easily performed on images. And it provides high perceptual transparency. When we embed the watermark by using LSB the quality of the image will not degrade. The main drawback of LSB technique is its poor robustness to common signal processing operations because by using this technique watermark can easily be destroyed by any signal processing attacks. It is not vulnerable to attacks and noise but it is very much imperceptible.

Some other algorithms of spatial domain watermarking are briefly discussed below:

ii. Additive Watermarking: The most straightforward method for embedding the watermark in spatial domain is to add pseudo random noise pattern to the intensity of image pixels. The noise signal is usually integers like (-1, 0, 1) or sometimes floating point numbers. To ensure that the watermark can be detected, the noise is generated by a key, such that the correlation between the numbers of different keys will be very low [5].

iii. SSM Modulation Based Technique: Spread-spectrum techniques are methods in which energy generated at one or more discrete frequencies is deliberately spread or distributed in time. SSM based watermarking algorithms embed information by linearly combining the host image with a small pseudo noise signal that is modulated by the embedded watermark.

iv. Texture mapping coding Technique: This method is useful in only those images which have some texture part in it. This method hides the watermark in the texture part of the image. This algorithm is only suitable for those areas with large number of arbitrary texture images (disadvantage) [3], and cannot be done automatically. This method hides data within the continuous random texture patterns of a picture.

v. *Patchwork Algorithm:* Patchwork is a data hiding technique developed by Bender et al and published on IBM Systems Journal, 1996 [6]. It is based on a pseudorandom, statistical model. Patchwork imperceptibly inserts a watermark with a particular statistic using a Gaussian distribution. A pseudo randomly selection of two patches is carried out where the first one is A and the second is B. Patch A image data is brightened where as that of patch B is darkened (for purposes of this illustration this is magnified).

vi. Correlation-Based Technique: In this technique, a pseudorandom noise (PN) pattern says W(x, y) is added to cover image I(x, y). Iw(x, y) = I(x, y) + k*W(x, y) Where K represent the gain factor, Iw represent watermarked image ant position x, y and I represent cover image. Here, if we increase the gain factor then although it increases the robustness of watermark but the quality of the watermarked image will decrease.

B. Frequency domain watermarking techniques:

Compared to spatial-domain methods, frequency-domain methods are more widely applied. The aim is to embed the watermarks in the spectral coefficients of the image. The most commonly used transforms are the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), the reason for watermarking in the frequency domain is that the characteristics of the human visual system (HVS) are better captured by the spectral coefficients [7]. Some of its main algorithms are discussed below:

i. Discrete cosine transforms (DCT): DCT like a Fourier Transform, it represents data in terms of frequency space rather than an amplitude space. This is useful because that corresponds more to the way humans perceive light, so that the part that are not perceived can be identified and thrown away. DCT based watermarking techniques are robust compared to spatial domain techniques. Such algorithms are robust against simple image processing operations like low pass filtering, brightness and contrast adjustment, blurring etc. However, they are difficult to implement and are computationally more expensive. At the same time they are weak against geometric attacks like rotation, scaling, cropping etc. DCT domain watermarking can be classified into Global DCT watermarking and Block based DCT watermarking. Embedding in the perceptually significant portion of the image has its own advantages because most compression schemes remove the perceptually insignificant portion of the image. Steps in DCT Block Based Watermarking Algorithm [8]

1) Segment the image into non-overlapping blocks of 8x8
2) Apply forward DCT to each of these blocks
3) Apply some block selection criteria (e.g. HVS)
4) Apply coefficient selection criteria (e.g. highest)
5) Embed watermark by modifying the selected coefficients.
6) Apply inverse DCT transform on each block.

ii. Discrete wavelet transforms (DWT): Wavelet Transform is a modern technique frequently used in digital image processing, compression, watermarking etc. The transforms are based on small waves, called wavelet, of varying frequency and limited duration. The wavelet transform decomposes the image into three spatial directions, i.e. horizontal, vertical and diagonal. Hence wavelets reflect the anisotropic properties of HVS more precisely. Magnitude of DWT coefficients is larger in the lowest bands (LL) at each level of decomposition and is smaller for other bands (HH, LH, and HL). The Discrete Wavelet Transform (DWT) is currently used in a wide variety of signal processing applications, such as in audio and video compression, removal of noise in audio, and the simulation of wireless antenna distribution. Wavelets have their energy concentrated in time and are well suited for the analysis of transient, time-varying signals. Since most of the real life signals encountered are time varying in nature, the Wavelet Transform suits many applications very well [9]. One of the main challenges of the watermarking problem is to achieve a better tradeoff between robustness and perceptivity. Robustness can be achieved by increasing the strength of the embedded watermark, but the visible distortion would be increased as well [9]. However, DWT is much preferred because it provides both a simultaneous spatial localization and a frequency spread of the watermark within the host image [10]. The basic idea of discrete wavelet transform in image process is to multi-differentiated decompose the image into sub-image of different spatial domain and independent frequencies [11].

Advantages of DWT over DCT:

Wavelet transform understands the HVS more closely than the DCT. Wavelet coded image is a multi-resolution description of image. Hence an image can be shown at different levels of resolution and can be sequentially processed from low resolution to high resolution. [2]

Disadvantages of DWT over DCT:

Computational complexity of DWT is more compared to DCT'. As Feig (1990) pointed out it only takes 54 multiplications to compute DCT for a block of 8x8, unlike wavelet calculation depends upon the length of the filter used, which is at least 1 multiplication per coefficient [2]

iii. Discrete Fourier transform (DFT):

Transforms a continuous function into its frequency components. It has robustness against geometric attacks like rotation, scaling, cropping, translation etc. DFT shows translation invariance. Spatial shifts in the image affects the phase representation of the image but not the magnitude representation, or circular shifts in the spatial domain don't affect the magnitude of the Fourier transform.

Advantages of DFT over DWT and DCT:

DFT is rotation, scaling and translation (RST) invariant. Hence it can be used to recover from geometric distortions, whereas the spatial domain, DCT and the DWT are not RST invariant and hence it is difficult to overcome from geometric distortions. [2]

Table below shows comparisons of different watermarking algorithms. [12][13]

Algorithm	Advantages	Disadvantages
LSB	1. Easy to implement and understand 2. Low degradation of image quality 3. High perceptual transparency.	1. It lacks basic robustness 2. Vulnerable to noise 3. Vulnerable to cropping, scaling.
Correlation	1. Gain factor can be increased resulting in increased robustness	1. Image quality gets decreased due to very high increase in gain factor.
Patchwork	1. High level of robustness against most type of attacks	1. It can hide only a very small amount of information.
Texture mapping coding	1. This method hides data within the continuous random texture patterns of a picture.	1. This algorithm is only suitable for those areas with large number of arbitrary texture images.
DCT	1. The watermark is embedded into the coefficients of the middle frequency, so the visibility of image will not get affected and the watermark will not be removed by any kind of attack.	1. Block wise DCT destroys the invariance properties of the system. 2. Certain higher frequency components tend to be suppressed during the quantization step.
DWT	1. Allows good localization both in time and spatial frequency domain 2. Higher compression ratio which is relevant to human perception.	1. Cost of computing may be higher. 2. Longer compression time. 3. Noise/blur near edges of images or video frames
DFT	1. DFT is rotation, scaling and translation (RST) invariant. Hence it can be used to recover from geometric distortions	1. Complex implementation 2. Cost of computing may be higher.

CONCLUSIONS

In this paper we have presented watermarking overview and also briefly discussed various watermarking techniques. Apart from this a brief and comparative analysis of watermarking techniques is presented with their advantages and disadvantages which can help in the new research areas.

REFERENCES

[1] Jiang Xuehua, ―Digital Watermarking and Its Application in Image Copyright Protection‖, 2010 International Conference on Intelligent Computation Technology and Automation.
[2] V. M. Potdar, S. Han and E. Chang, "A Survey of Digital Image Watermarking Techniques", 2005 3rd IEEE International Conference on Industrial Informatics (INDIN).
[3] N. Chandrakar and J. Baggaa,"Performance Comparison of Digital Image Watermarking Techniques: A Survey", International Journal of computer Application Technology and Research, vol. 2, no. 2, (2013), pp. 126-130.
[4] D. Mistry," Comparison of Digital Watermarking Methods" (IJCSE) International Journal on Computer Science and Engineering, vol. 02, no. 09, (2010), pp. 2805-2909.
[5] CHAPTER 2: LITERATURE REVIEW, Source: Internet
[6] http://ippr-practical.blogspot.in
[7] Manpreet kaur, Sonia Jindal, Sunny behal, ―A Study of Digital image watermarking‖, Volume2, Issue 2, Feb 2012.
[8] Vidyasagar M. Potdar, Song Han, Elizabeth Chang, ―A Survey of Digital Image Watermarking Techniques‖, 2005 3rd IEEE International conference on Industrial Informatics (INDIN).
[9] Evelyn Brannock, Michael Weeks, Robert Harrison, Computer Science Department Georgia State University ―Watermarking with Wavelets: Simplicity Leads to Robustness‖, Southeast on, IEEE, pages 587 – 592, 3-6 April 2008.
[10] G. Bouridane. A, M. K. Ibrahim, ―Digital Image Watermarking Using Balanced Multi wavelets‖, IEEE Transaction on Signal Processing 54(4), (2006), pp. 1519-1536.
[11] Cox, I.J.; Miller, M.L.; Bloom, J.A., ―Digital Watermarking,‖ Morgan Kaufmann, 2001.
[12] Jiang Xuehua, ―Digital Watermarking and Its Application in Image Copyright Protection‖, 2010 International Conference on Intelligent Computation Technology and Automation.
[13] Amit Kumar Singh, Nomit Sharma, Mayank Dave, Anand Mohan, ―A Novel Technique for Digital Image Watermarking in Spatial Domain‖, 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing.

About Author
Aaqib Rashid is a Research Scholar and has completed MCA from Kashmir University and M.Phil in Computer Science from Dr. C.V Raman University India.

International Journal of Computer Applications Technology and Research
Volume 5–Issue 3, 151-155, 2016, ISSN:2319–8656

Methodology of Implementing the Pulse code techniques for Distributed Optical Fiber Sensors by using FPGA: Cyclic Simplex Coding

Yelkal Mulualem
Lecturer, Department of Information Technology,
College of Natural and Computational Science,
University of Gondar, Ethiopia.

ABSTRACT: In recent researches Coding techniques are used in OTDR approach improve Signal-to-Noise Ratio (SNR). For example, the use of simplex coding (S-coding) in conjunction with OTDR can be effectively used to enhance the Signal-to-Noise Ratio (SNR) of the backscattered detected light without sacrificing the spatial resolution; In particular, simplex codes have been demonstrated to be the most efficient among other suitable coding techniques, allowing for a good improvement in SNR even at short code lengths. Coding techniques based on Simplex or Golay codes exploit a set of different sequences (i.e. codes) of short (about 10 ns) NRZ laser pulses to increase the launched energy without impairing the spatial resolution using longer pulse width. However, the required high repetition rate of the laser pulses, hundreds of MHz for meter-scale spatial resolution, is not achievable by high peak power lasers, such as rare-earth doped fibre or passive Q-switched ones, which feature a maximum repetition rate of few hundred kHz. New coding technique, cyclic simplex coding (a subclass of simplex coding), tailored to high-power pulsed lasers has been proposed. The basic idea is to periodically sense the probing fibre with a multi-pulse pattern, the repetition period of which is equal to the fibre round-trip time. This way, the pattern results as a code spread along the whole fibre, with a bit time inversely proportional to the code length. The pulse width can be kept in the order of 10 ns to guarantee a meter-scale spatial resolution and the peak power can be set close to the nonlinear effect threshold.

Keywords: Signal, Noise, Ratio, Laser and Pulse.

I. INTRODUCTION

The purpose of this research work is to present the design and analysis of new FPGA architectures, aiming to address the main design issues related to Decoding of the averaged Stoke and Anti-Stoke traces. The main task performed by the new architecture implemented on the FPGA is to decode Stokes and anti-Stokes trace samples coming from the ADC[1][2]. The whole FPGA architecture has been developed using the Verilog hardware description language. This new FPGA architectures has three different sub-modules. Each module varies with respect to each other in terms of their functionality. The goal of the analysis is to develop FPGA architecture that can be able to decode averaged Stock and Anti-Stoke traces using minimum resource utilization. The following list is an overview of the Top-module and its three sub-module architecture considered.

A. Top Module:

The total operation of the system is performed in a single clock cycle. It is referred to as Top Module because it is the outer interface interacting with FPGA board. It takes averaged coded Stoke and Anti-Stoke trace data and codeword bit pattern, and then it returns the decoded Stoke and Anti-Stoke sampled trace data.

B. Read-Codeword Module:

All reading operations are performed in a single clock cycle. It is referred to as Read-Codeword Module because we have code word bits stored in the register. It takes code patterns and returns the code words bit by bit.

C. Read-RAM Module:

All reading operations are performed in a single clock cycle. It is referred to as Read RAM Module because there is averaged coded Stoke and Anti-Stoke sampled traces stored in the Dual port RAM. It reads the averaged Stoke and Anti-Stoke sampled traces from the Dual port RAM and returns one averaged sample per clock cycle.

D. Decoder Module

All Decoding operations are performed in a single clock cycle. It is referred to as Decoder Module because it decodes averaged coded Stoke and Anti-Stoke sampled traces. It takes both single bit codeword and single averaged Coded Stoke and Anti-Stoke sampled trace data and returns the Decoded Stoke and Anti Stoke trace data.

The contributions made by this research work include a synthesizable Verilog description of each of the module architectures described above, a synthesizable top module Verilog interface between the FPGA and the development platform used for this research.

The energy of the probing laser pulse cannot be freely increased. The energy of the launched pulse is indeed bounded by the targeted spatial resolution, which implies a small pulse width, and by the threshold for the onset of the fibre nonlinearities, which upper bounds the pulse peak power level. [11]

II. RELATED WORK

Pulse coding is the typical solution adopted to address the issues of averaging. Its basic principle is, launching proper laser pulse sequences instead of a single pulse, so as to increase the probing energy without impairing the spatial resolution.
These sequences are the optical representation of binary linear algebraic codes, which are widely used in communication theory for error detection and correction. [3]

Different codes families are grouped in to classes, each of them containing codes of the same length. Once a code class of length M is selected, a code set of M codes is built. Then, each code of the set is launched, and its Stokes and anti-Stokes responses are acquired. Finally, the set of responses is decoded to obtain a couple of Stokes and Anti-Stokes traces to be used for the temperature assessment. As described in the above, the most important aspect of pulse coding is that the SNR of decoded traces increases with M.

In particular, it had been shown that for such applications Simplex codes provide the best performance in terms of coding gain, i.e. for a given M they allow to achieve the best SNR enhancement with respect to other coding schemes [3]. It is possible to build a simplex code set for any M = 4n + 1, with n = 1, 2,3,4,....In Figure 1.1 a qualitative example for M = 3 is reported[3][4][5].

The code set is [011], [101], [110]. Whenever a laser pulse is launched, i.e. whenever the code bit is 1, a new backscattered trace starts. This means that the response R (t) to the code c acquired by the receiver is given by the overlapping of some delayed replicas of the trace (t) to be recovered [7][8][9]. The delay is a multiple of the chosen code bit time. In the reported example, the code responses are given by [10][11],

$$R_{011}(t) = \psi(t-\tau) + \psi(t-2\tau)$$
$$R_{101}(t) = \psi(t) + \psi(t-2\tau)$$
$$R_{101}(t) = \psi(t) + \psi(t-\tau)$$

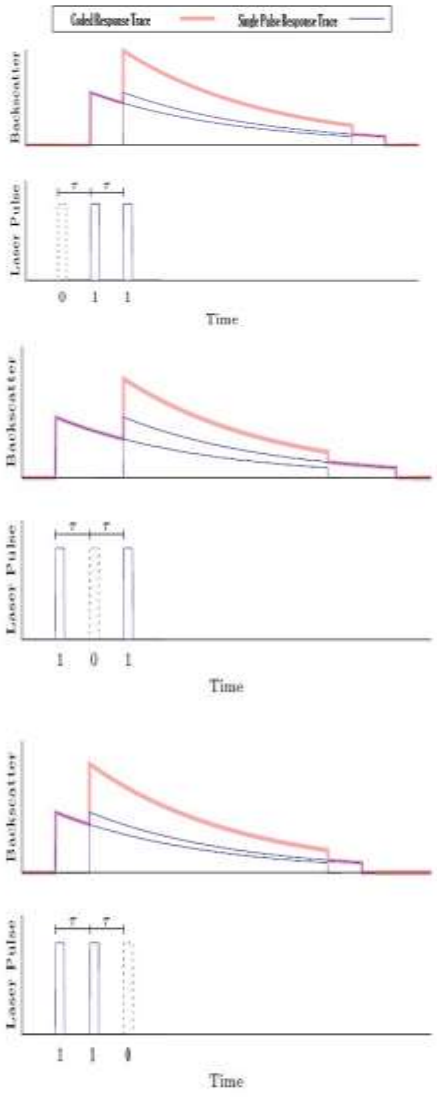


Figure 1.1. Example for Conventional Simplex Coding M = 3

Rearranging the above equations in matrix form, it follows,

$$\begin{bmatrix} R_{011}(t) \\ R_{101}(t) \\ R_{111}(t) \end{bmatrix} = S \begin{bmatrix} \psi(t) \\ \psi(t-\tau) \\ \psi(t-2\tau) \end{bmatrix} with S = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \text{.... [1]}$$

III. IMPLEMENTATION

In order to evaluate the performance of FPGA based implementation, the algorithm was coded in Verilog hardware description language and implemented on Virtex4 (family using Xilinx ISE 10.0.3 tool). To test the new developed FPGA architecture and to see the real SNR improvement provided by the cyclic coding, measurements with coding are compared to the ones obtained by the conventional technique using the same acquisition time and the same peak

power, the whole DTS system has been configured. The simulation have been performed by allocating 71 bits of code words along 26 km of SMF and the repetition rate of the laser has been set to constant value. Then, extraction of the averaged Coded Stokes and Anti-Stokes trace data after 100k acquisitions (100 time-averaged traces) has been done using Lab VIEW software. Figure 1.2 shows experimental setup to acquire the coded Stokes and Anti-Stokes traces.

Figure 1.2. Experimental Setup to Acquire Coded Stokes and Anti-Stokes Traces

As described in the chapter two, Multi pulse pattern can be obtained by triggering the laser at the fixed rate and by implementing the cyclic code through an external modulator (acousto-optic Modulator), which allows to filter out a pulse if the corresponding bit code is equal to 0. This way, the pulsed laser operates at a constant frequency, which guarantees a good repeatability of the generated pulse shape and peak power.

Figure 1.3 and 1.4 shows diagram of acquired coded Stokes and anti-Stokes trace. Most of the trace recovery blocks have been implemented using Lab VIEW and others software's the likes of Microsoft-Excel

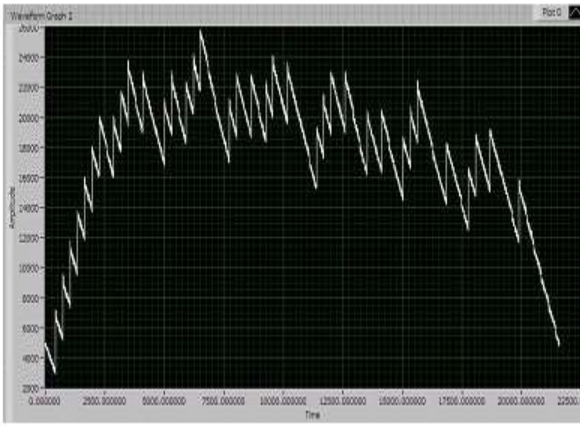


Figure 1.3. Acquired Waveform of Averaged Coded Stokes Trace, with 71 bit cyclic Simplex codes

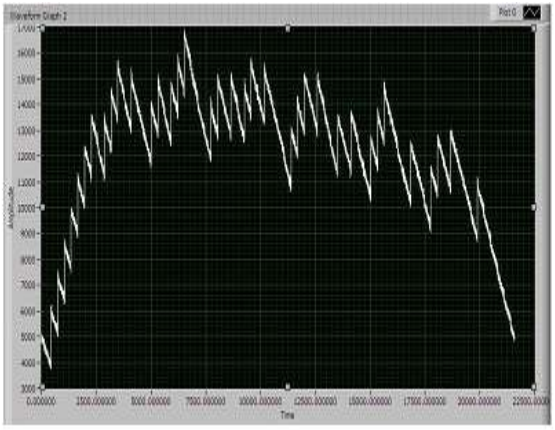


Figure 1.4. Acquired Waveform of Averaged Coded Anti-Stokes Trace, with 71 bit cyclic Simplex codes

The new FPGA architecture modules have been developed using the Verilog hardware description language. Each step of the design flow, i.e. the logical synthesis thesis, the functional simulations, the implementation and the final post place and route simulations; have been carried out within the Xilinx Integrated Software Environment (ISE) 10.1.03. Test Bench architecture that has been implemented to test multi-pulse patterns in newly developed FPGA architecture modules were carried out using Xilinx integrated Software Environment (ISE) 10.1.03. The simulation has been done with code word pattern length of 71 bit, Number of Samples per slot(number of samples for each single pulse trace in a multi-pulse technique) is 304, total number of samples (multiple of code word pattern length and Number of Samples per slot) 21584 and with working clk-frequency of 150 MHz (6 ns). Figure 1.5, 1.6 and 1.7 shows the simulation waveforms of each modules of new developed FPGA architecture.

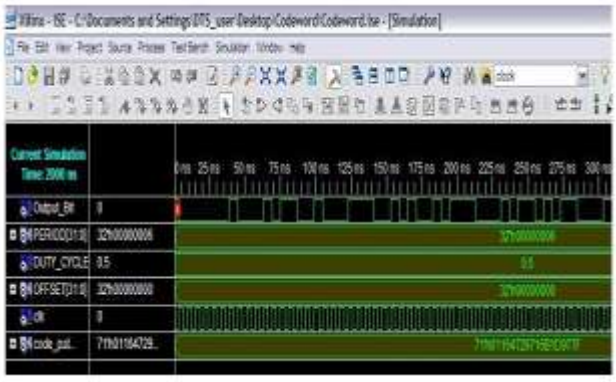


Figure 1.5 Simulation Waveform of Read-Codeword Module

Figure 1.6 Simulation Waveform of Read-RAM Module

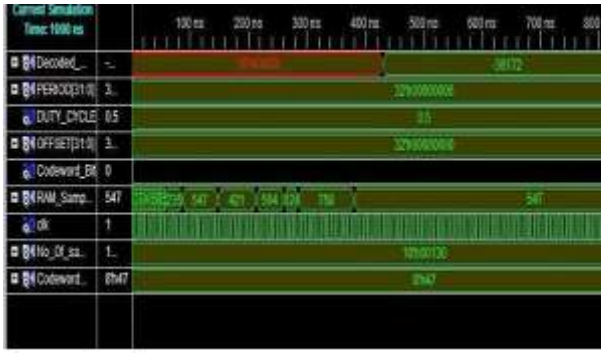


Figure 1.7 Simulation Waveform of Decoder Module

Figure 1.5, shows Simulation Waveform of Read-Codeword Module, which com- prises input of codeword bit patterns (e.g. 71 codeword bit pattern) and single bit output called output-bit.

In Figure 1.6, Simulation Waveform of Read-RAM Module is shown. It has 3 input called Number codeword bit pattern which is defined in the design phase of this thesis report, Number of Sample per Slot, (e.g 304) and clock frequency (clk). It has one output called Sample Data-out which will serve as an input for Decoder Module.

Figure 1.7, shows Simulation Waveform of Decoder Module which is the core of the new developed FPGA architecture. It takes an input from the output of the other two modules, Read-Codeword Module and Read-RAM Module. Finally it returns an output called Decoded Data, final decoded data of Averaged coded stoke and anti-stoke trace data. As a common all the three modules have an input called clock frequency (clk) which is 150 MHz (6 ns).

In figure 1.8, the utilization statistics of the Top Module (i.e. main FPGA logic resources) are shown. It can be noted that the IO Blocks are the most used which is 18%, whereas only the 2% of the available slices are occupied. This means that there is still a great room for the future development of other functionalities.

Top_Module Partition Summary
No partition information was found.

Device Utilization Summary (estimated values)			
Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slices	152	5472	2%
Number of Slice Flip Flops	144	10944	1%
Number of 4 input LUTs	305	10944	2%
Number of bonded IOBs	45	240	18%
Number of GCLKs	1	32	3%

Figure 1.8 FPGA Device Utilization Summary

IV. CONCLUSION

The main issues related to the design of new FPGA architecture to decode the averaged coded Stokes and Anti-Stokes multi-pulse traces data have been analyzed and addressed. This research is a small part of a large project which is undergoing at the TECIP labs of School Superior Sant'Anna. The primary focus has been on the design and implementation of decoding algorithmic module on FPGA. After describing the basic working principles of Distributed Temperature Sensor (DTS) systems, an overview of FPGA(Field Programmable Gate Array) and Verilog High level Description language has been provided, and also a new FPGA architecture which used to decode averaged coded Stokes and Anti-Stokes traces has been designed and examined in detail. Finally, the trade o s between their performance parameters has been analyzed. The new developed FPGA architecture to decode averaged coded Stoke and Anti-Stoke traces has been experimentally demonstrated and preliminary test results showed the expected performance confirming the validity of the result of embedded FPGA decoding. It lets to DTS systems an embedded decoding with less PC resources and less Overhead time for the communication to the PC. In general, laser pulse coding technique allows to significantly improving the SNR of the acquired traces. This advanced coding technique based on cyclic coding is well suited for long range DTS systems, as it can easily be implemented also with high power pulsed lasers featuring limited repetition rates.

V. REFERENCES

[1]. A. H. Hartog, "Progress in distributed fiber optic temperature sensing," in Proc. of Fiber Optic Sensor Technology and Applications Conference 2001, M. A. Marcus and B. Culshaw, Eds., vol. 4578. SPIE, 2002, pp. 4352.

[2]. D. A. Krohn, "Fiber Optical Sensors, Fundamentals and Applications, 3rd Ed" Research Triangle Park, NC, Instrument Society of America, 2000.

[3]. F. Di Pasquale "SNR enhancement of Raman based long-range distributed temperature sensors using cyclic Simplex codes" F. Baronti, A. Lazzeri, R.Roncella, R. Saletti A. Signorini, M. A. Soto, G. Bolognini,

[4]. http: ==en:wikipedia:org=wiki=Distributed Temperature Sensing.

[5]. J. M. Lopez-Higuera Ed., Handbook of Optical Fiber Sensing Technology, Chichester,U.K., John Wiley and Sons Ltd., 2001.

[6]. A. Rogers, Distributed optical fiber sensing, in Handbook of Optical Fiber Sensing Technology", J. M. Lopez-Higuera Ed. Chichester, U.K., John Wiley and Sons Ltd., 2001, ch. 14, p.271-312.

[7]. J. P. Dakin, Distributed optical fiber systems, in Optical Fiber Sensors: systems and Applications", B. Culshaw, J. Dakin, Eds. Norwood, MA, Artech House, 1988, vol.2, ch. 15, p. 575-598.

[8]. S. Adachi, "Distributed optical fiber sensors and their applications," in Proc.SICE Annual Conference, 2008, pp. 329333.

[9]. A. H. Hartog, A distributed temperature sensor based on liquid-core optical fibers, J. Lightwave Technol. 1(3), pp. 498 - 509, 1983.

[10]. S. V. Shatalin, V. N. Treschikov, A. J. Rogers, Interferometric Optical Time- Domain Reectometry for Distributed Optical-Fiber Sensing, Applied Optics.

[11]. "Raman-based distributed temperature sensor with 1m spatial resolution over 26km SMF using low-

repetition-rate cyclic pulse coding " Marcelo A. Soto,1 Tiziano Nannipieri,1 Alessandro Signorini,1 Andrea Lazzeri,2 Federico Baronti,2 Roberto Roncella,2 Gabriele Bolognini,1,* and Fabrizio Di Pasquale1.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 156 - 158, 2016, ISSN:- 2319–8656

A Review on Feature Extraction Techniques and General Approach for Face Recognition

Aakanksha Agrawal
Department of CSE
RCET, Bhilai
CG, India

Steve Samson
Department of CSE
RCET, Bhilai
CG, India

Abstract: In recent time, alongwith the advances and new inventions in science and technology, fraud people and identity thieves are also becoming smarter by finding new ways to fool the authorization and authentication process. So, there is a strong need of efficient face recognition process or computer systems capable of recognizing faces of authenticated persons. One way to make face recognition efficient is by extracting features of faces. Several feature extraction techniques are available such as template based, appearance-based, geometry based, color segmentation based, etc. This paper presents an overview of various feature extraction techniques followed in different reasearches for face recognition in the field of digital image processing and gives an approach for using these feature extraction techniques for efficient face recognition.

Keywords: face recognition, feature extraction, lip extraction, eye extraction

1. INTRODUCTION

Feature Extraction is a kind of process for reducing dimensionality so as to represent the interesting image parts with great efficiency as a compact feature vector [16]. This is useful in case of large images.

No one can say which algorithm is best suitable for extracting features. The algorithm selection is dependent on: (1) What exactly is the task it needs to perform (2) Whether supervised method is needed or unsupervised (3) Whether Inexpensive method is required or strong computational method is required etc. Some techniques for feature extraction are Speeded Up Robust Features (SURF), Histogram of Oriented Gradients (HOG), Color Histograms, Local Binary Patterns (LBP), Haar wavelets, etc. [16]

Many software packages for data analysis are available providing feature extraction and dimension reduction. Some numerical programming environments such as MATLAB, NumPy, SciLab also provides technique for feature extraction like Principal Component analysis (PCA) etc. [18]

2. FACE FEATURES FOR EXTRACTION

2.1 Lip Feature

Lip is a sensory organ existing in visible portion of human face and considered to be different for each and every individual. There are researches carried out for face recognition and classification of gender using lip shape and color analysis. [3]

2.2 Eye/Iris Feature

Eye/Iris has randomness due to its small tissues which provides differentiation to the pattern of eye for each and every individual human being. [13] The stableness, uniqueness and non-invasion, these qualities make the iris outstanding among several biometric features. [14]

2.3 Nose Feature

The nose tip is a distinctive point of human face. It also remains unaffected even due to changes in facial expressions. [18] Thus, it proves to be efficient for face recognition.

3. TECHNIQUES FOR FEATURE EXTRACTION PURPOSE

3.1 Face Part Detection (FPD)

Convolution technique is used in this algorithm. It works by multiplying vectors and returns values by using length and width. Gabor features are done using Gabor filters and here image decomposition is done by converting real part and imaginary part. [12]

3.2 Principal Component Analysis (PCA)

In this, the dimensionality is reduced by projecting the data onto the largest eigenvectors. [18] It selects the weights on the basis of frequency in the frequency domain. It cannot separate the class linearly.

3.3 Linear Discriminant Analysis (LDA)

LDA is a generalization of Fisher's linear discriminant to find a linear combination of features that characterizes two or more classes of events. The resulting combination may be used for dimensionality Reduction or as a linear classifier. [18]

3.4 Speeded Up Robust Features (SURF)

Surf algorithm is an improvement of Scale Invariant Feature Transform (SIFT) algorithm. Surf uses a fast multi-scale Hessian keypoint detector that can find keypoints. It can also be used to compute user specified keypoints. Only 8 bit grayscale images are supported. [3]

4. RELATED WORK

4.1 Rutuja G. Shelke, S.A.Annadate presented a novel approach for Face Recognition and Gender classification strategy using the features of lips. Here feature extraction is carried out by using Principal component analysis (PCA) and Gabor wavelet. Out of two techniques, results of Gabor filter are more accurate and fast because it is having less leakage in time frequency domain. [International Journal of Innovation and Scientific Research (IJISR), Vol 10, No.2, Oct.2014, Innovative Space of Scientific Research Journals (ISSR)].

4.2 Ishvari S. Patel, Apurva A. Desai used Preprocessing techniques like Edge Detection by Canny Method and Height and Width comparison for Lip Contour Detection. This model works effectively and gives around 98% result for image sequences but we can still improve accuracy of result by extracting perfect lips region. [International Journal of Scientific Research (IJSR), Volume II, Issue V, May 2013].

4.3 Sambit Bakshi, Rahul Raman, Pankaj K Sa paper proposes that grayscale lip images constitute local features. The claim has been experimentally established by extracting local features applying two techniques viz. SIFT and SURF. The results obtained are enough to establish that unique local features exist in lip images through which an individual can be recognized. [India Conference (INDICON), 2011 Annual IEEE].

4.4 Sasikumar Gurumurty, B. K. Tripathy divided methodology into: Mouth Region Localization and Key point's Extraction and Model Fitting. In first and second steps**,** mouth region and key points are found by using hybrid edges**,** which combine color and intensity information. In third step**,** cubic polynomial models are fitted using key points position and hybrid edges. An automatic, robust and accurate lip segmentation method has been presented. This is considered as good result and encourage for its use combined with other biometrics systems.[I.J. Intelligent Systems and Applications (IJISA), July 2012 in MECS].

4.5 B. Sangeetha Devi, V.J.Arul Karthick used two processes for lip recognition. First, face detection by Viola and Jones algorithm. Second, lip detection by morphological operation and five various mouth corner points. Lip biometric can be used to authenticate an individual since the lip is unique. [International Journal of Advanced Research Trends in Engineering and Technology (IJARTET), Vol. II, Special Issue I, March 2015].

5. GENERAL APPROACH FOR FACE RECOGNITION

5.1 Acquiring the image of an individual's face

Digitally scan an existing photograph, or Acquire a live picture of a subject.

5.2 Locate image of face

Software is used to locate the faces in the image that has been obtained.

5.3 Analysis of facial image

Software measures face according to its peaks and valleys and focuses on the inner area of the face.

5.4 Comparison

The face print created by the software is compared to all face prints the system has stored in its database.

5.5 Match or No Match

Software decides whether or not any comparisons from the above step are close enough to declare a possible match.

6. PERFORMANCE EVALUATION PARAMETERS

6.1 False Acceptance Rate (FAR)

The probability that a system will incorrectly identify an individual or will fail to reject an imposter.
(Also called Type2 Error Rate)

FAR=NFA/NIIA

NFA=number of false acceptance
NIIA=number of imposter identification attempts

6.2 False Rejection Rate (FRR)

The probability that a system will fail to identify an enrolee.
(Also called Type1 Error Rate)

FRR=NFR/NEIA

NFA=number of false rejection
NIIA=number of enrolee identification attempts

7. SCOPE OF FUTURE WORK

Face Recognition is a very vast and elaborated field. It has no end. As the advancement in science and technology, new techniques will continue developing day-by-day. Today, Lip and Eye extraction are mostly discussed techniques for face recognition purpose but in future many more advance techniques will arise for performing Face Recognition with much more accuracy and efficiency.

8. ACKNOWLEDGMENTS

Our sincere thanks to all the respected and experienced faculties for their valuable guidance and motivation that always encouraged us to give our full dedication towards a new improvement in the field of science and image processing.

9. REFERENCES

[1] Rutuja G. Shelke and S. A. Annadate, "Face Recognition and Gender Classification Using Feature of Lips", International Journal of Innovation and Scientific Research (IJISR), Innovative Space of Scientific Research Journals (ISSR), Vol. 10, No. 2, Oct. 2014.

[2] B. Sangeetha Devi and V. J. Arul Karthick, "Lip Recognition With Support Vector Machine (SVM) Classifier", International Journal of Advanced Research Trends in Engineering and Technology (IJARTET), Vol. II, Special Issue I, March 2015.

[3] Sambit Bakshi, Rahul Raman, Pankaj K Sa, "Lip pattern recognition based on local feature extraction", India Conference (INDICON), IEEE Annual, 2011.

[4] Sunil Sangve, Nilakshi Mule, "Lip Recognition for Authentication and Security", IOSR Journal of Computer Engineering (IOSR-JCE) Volume 16, Issue 3, Ver. VII, May-Jun. 2014.

[5] Ishvari S. Patel and Apurva A. Desai, "Lip Segmentation Based on Edge Detection Technique", International Journal of Scientific Research (IJSR), Volume II, Issue V, May 2013.

[6] Duy Nguyen, David Halupka, Parham Aarabi and Ali Sheikholeslami, "Real-Time Face Detection and Lip Feature Extraction Using Field-Programmable Gate Arrays", IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 36, No.4, August 2006.

[7] Sasikumar Gurumurty and B. K. Tripathy, "Design and Implementation of Face Recognition System in Matlab Using the Features of Lips", I.J. Intelligent Systems and Applications (IJISA), July 2012 in MECS.

[8] Jyoti Bedre and Shubhangi Sapkal, "Comparative Study of Face Recognition Techniques", Emerging Trends in Computer Science and Information Technology (ETCSIT2012), International Journal of Computer Applications (IJCA), 2012.

[9] Riddhi Patel and Shruti B. Yagnik, "A Literature Survey on Face Recognition Techniques", International Journal of Computer Trends and Technology (IJCTT), Volume 5, No.4, Nov 2013.

[10] John Daugman, "How Iris Recognition works", IEEE Transactions on Circuits and Systems for Video Technology, Volume. 14, No. 1, Jan 2004.

[11] Mayank Vatsa, Richa Singh, and Afzel Noore, "Improving Iris Recognition Performance Using Segmentation, Quality Enhancement, Match Score Fusion, and Indexing", IEEE Transactions on Systems, Man and Cybernetics— Part B: Cybernetics, Feb 2008.

[12] Dr. S. Vijayarani, S. Priyatharsini, "Facial Feature Extraction Based On FPD and GLCM Algorithms", International Journal of Innovative Research in Computer and Communication Engineering, Vol.3, Issue 3, March 2015.

[13] Pankaj K.Sa, S.S. Barpanda, Bansidhar Manjhi, "Region Based Feature Extraction from Non-Cooperative Iris Images", Innovation Syst Software Engg, 2015.

[14] Changcheng Li, Weidong Zhou, Shasha Yuan, "Iris Recognition Based on a Novel Variation of Local Binary Pattern", Springer-Verlag Berlin Hiedelberg, Vis Comput, 2014.

[15] Rafael C. Gonzalez, Richard E. Woods and Steven L. Eddins, "Digital Image Processing Using MATLAB", Second Edition

[16] The MathWorks, Inc., "Image Processing Toolbox", User's Guide, COPYRIGHT 1993–2015

[17] S. N. Sivanandam, S. Sumathi, S. N. Deepa "Neural Network using MATLAB 6.0".

[18] www.google.com

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 159 - 164, 2016, ISSN:- 2319–8656

Solve Big Data Security Issues

Abhinandan Banik
IBM India Pvt. Ltd.

Samir Kumar Bandyopadhyay
Department of Computer Science and Engineering,
University of Calcutta
Kolkata, India

Abstract: The advent of Big Data has presented nee challenges in terms of Data Security. There is an increasing need of research in technologies that can handle the vast volume of Data and make it secure efficiently. Current Technologies for securing data are slow when applied to huge amounts of data. This paper discusses security aspect of Big Data.

Keywords: Big Data; Challenges; Security and Privacy

1. INTRODUCTION

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The concept of big data has been endemic within computer science since the earliest days of computing. "Big Data" originally meant the volume of data that could not be processed by traditional database methods and tools. Each time anew storage medium was invented, the amount of data an accessible exploded because it could be easily accessed.

Big Data is characterized by three aspects: (a) the data are numerous, (b) the data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed very quickly. Big Data is promising for business application and is rapidly increasing as a segment of the IT industry. The variety, velocity and volume of big data amplifies security management challenges that are addressed in traditional security management. Big data security challenges arise because of incremental differences, not fundamental ones. The differences between big data environments and traditional data environments include:

- The data collected, aggregated, and analyzed for big data analysis
- The infrastructure used to store and house big data
- The technologies applied to analyze structured and unstructured big data

The variety, velocity and volume of big data amplifies security management challenges that are addressed in traditional security management. Big data repositories will likely include information deposited by various sources across the enterprise. This variety of data makes secure access management a challenge. Each data source will likely have its own access restrictions and security policies, making it difficult to balance appropriate security for all data sources with the need to aggregate and extract meaning from the data.

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. The 3Vs that define Big Data are Variety, Velocity and Volume.

1) Volume: There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.
2) Velocity: Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.
3) Variety: Today, data comes in all types of formats. Structured, numeric data in traditional databases.

Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. We need to find ways of governing, merging and managing these diverse forms of data.

As computer applications were developed to handle financial and personal data, the need for security is necessary. The data on computers is an extremely important aspect for processing and transmitting secure data in various applications. Security is the process of preventing and detecting unauthorized use of computer or network.

Prevention measures help us to stop unauthorized users

from accessing any part of computer system. Detection helps to determine whether or not someone attempted to break into the system. The goal of cryptography is to make it possible for two people can exchange a message in such a way that other people cannot understand the message. There is no end to the number of ways this can be done, but here we will be concerned with methods of altering the text in such a way that the recipient can undo the alteration and discover the original text.

2. REVIEW WORKS

Big data is a large set of unstructured data even more than tera and peta bytes of data. Big Data[1] can be of only digital one. Data Analysis become more complicated because of their increased amount of unstructured or semi-structured data set. Predictions, analysis, requirements etc., are the main things that should be done using the unstructured big data. Big data is a combination of three v's those are namely Volume, Velocity and Variety. Big data will basically processed by the powerful computer. But due to some scalable properties of the computer, the processing gets limited.

For encryption/decryption process, in modern days is considered of two types of algorithms viz., Symmetric key cryptography and Asymmetric key cryptography.

Symmetric key cryptography:
Symmetric-key algorithms are those algorithms that use the same key for both encryption and decryption. Examples of symmetric key algorithms are Data Encryption Standard (DES) and Advanced Encryption Standard (AES).

Asymmetric key cryptography:
Asymmetric-key algorithms are those algorithms that use different keys for encryption and decryption. Examples of asymmetric-key algorithm are Rivest- Shamir-Adleman (RSA) and Elliptic curve cryptography (ECC).

Big data deals with storing the data, processing the data, retrieval of data. Many technologies are used for these purposes just like memory management, transaction management, virtualization and networking. Hence security issues of these technologies are also applicable for big data.

3. PROPOSED METHOD

In a symmetric cryptosystem model we use the same key for encryption and decryption. In this proposed model our main key has two parts. The first part of the main key is substitution key or sub-key and the last part is shuffling key or suf-key. Here last 10bits of main key represents suf-key and apart from that last 10bits rest of the part represents sub-key. We have to remember that there is a restriction of choosing sub-key; it can only bean 8-bit or a multiple of 8 i.e.16, 24, 32 and so on. So, main key of our proposed Big data have many security issues and these issues can be solved using some approaches like data encryption.

Data Encryption Standards or DES is a block encryption algorithm. When 64-bit blocks of plain text go in, 64-bitblocks of cipher text come out. It is asymmetric algorithm, meaning the same key is used for encryption and decryption. It uses a 64-bitkey, 56 bits make up the true key, and 8 bits are used for parity. When the DES algorithm

is applied to data, it divides the message into blocks and operates on them one at a time. A block is made up of 64bits and is divided in half and each character is encrypted one at a time. The characters are put through 16 rounds of transposition and substitution functions. The order and type of transposition and substitution functions depend on the value of the key that is inputted into the algorithm. The result is a 64-bitblock of cipher text. There are several modes of operations when using block ciphers. Each mode specifies how a block cipher will operate. One mode may work better in one type of environment for specific functionality, whereas another mode may work in a different environment with totally different types of requirements.

The Advanced Encryption Standard or AES [6] algorithm is also a symmetric block cipher that can encrypt and decrypt information. This algorithm is capable of using cryptographic keys of 128,192and 256 bits to encrypt and decrypt data in blocks of 128 bits. The input and output for the AES algorithm each consist of sequences of 128bits(digits with values of 0 or 1).These sequences will sometimes be referred to as blocks and the number of bits they contain will be referred to as their length. The bits within such sequences will be numbered starting at zero and ending at one less than the sequence length (block length or key length).The number i attached to a bit is known as its index and will be in one of the ranges between 0 to127or 0 to255 or 0 to 255 depending on the block length and key length.

In this paper, we concentrate on developing an innovative cryptosystem for information security which is more advantages than existing symmetric key standards like DES and AES in context of security.

model will always be sub-key +10bits,i.e.18, 26, 34,42 and so on. Therefore this proposed model is based on substitution-expansion-shuffling technique.

In this method first determine the sub-key from the main key. On the basis of this sub-key, first step, method of substitution will occur. Let take an 8-bit suf-key ,i.e. 00111010.Firstcalculate the key value to perform the substitution. In case of 8-bit key this values can be

determine by calculating the decimal value of first three bits then next and so on. In our key 1st three bits i.e. 0,1and 2 positions are 001whose corresponding decimal value is 1.Consider first value. Then calculate next three bits i.e.1,2 and 3 three bit positions i.e. 011 is 3. Consider second value. Perform same technique until come to last bit positioni.e.7^{th} position .Now, what happen if we try to take next three bits from the 6th position where only 2-bits left or in case of 7th position where only 1-bitleft? In such case we take rest of the bits from the beginning e.g. for 6^{th} position it should be 100 and for 7^{th} position it is 000.In case of 8-bits sub-key we'll get eight key values lie in between 0 to7.Key values of our sub-key are1, 3, 7, 6,5,2,4 and 0.As soon as we'll get that key values we create a block of eight rows and 1+Ncolumn (N is number of characters in the file). Here first column represent the key values of the sub-key and remaining columns represent the characters of the file. Now, we convert each character to its corresponding 8-bit binary stream(as we take 8-bit sub-key)and put that into the block serially from top to bottom. After the whole block is filled up it will rearrange in ascending order from top to bottom on the basis of the key values and thus the bits of the whole plain text will have substituted and we will get the substituted binary bit stream. Though we have to determine that key values therefore the number of sub-key combinations we can generate by a particular sub-key length is$2^{(\text{Length of sub-key}/2)}$,i.e.for8-bit we can generate up to 2^4 combinations of sub-key,for16-bit we can generate 2^8combinations of sub key and so on. Now, key values are varies with the sub-keys, like in case of 16-bit key it lies between 0 to15, for24-bit key it lies between 0-23 and so on. Now we convert the characters to corresponding binary stream of that bit which our sub-key has, i.e. incase of16-bit key we convert the character to 16-bit binary stream, for24-bit key we convert to 24-bit binary stream and so on.

The next step after method of substitution is method of expansion. As the name suggested we can guess that there we will perform some sort of expansion. This expansion will perform on the substituted binary bit stream and it will also need the key values of the sub-key. Now, key value of our sub-key is1,3,7,6,5,2,4 and 0. At the beginning we take the first key value which is1. Now ,from the substituted binary bit stream we will take 1^{st} bit because our key value is1 and we also perform only one time expansion as our value is1.We expand up to 8 bit in case of 8-bit sub-key. As we have 1^{st} bit 1 so we add another 7 bits i.e.0101010.0101010 because last bit is1, thus we add alternate 0 and 1.We then move to next key value which is 3,therefore we perform up to three times expansion. But we perform expansion from the starting key value (1^{st} key value to 3^{rd} key value) i.e.1,3 and7. We follow this technique until the substituted binary stream found empty. As soon as we found it is empty,we will stop expansion and go to next step.We will expand up to8-bit to make expanded bit stream divisible by 8.It helps to convert binary bit stream to corresponding cipher text at the end. Apart from8-bit substitution key we follow quite different technique for expansion. The basic idea is same with only one change. For rest of the sub-keys, if key values are less than the half of the length of current sub-key we will expand it up to the length of previous sub-key and if key values are greater than the half of the length of current sub-key we will expand it up to the length of current sub-key. E.g. in case of 16 bit sub-key ifkeyvalueis6thenwewillexpandup to 8 bit, but if key value is 10 then we will expand up to 16 bit. Now, it can be a situation that our key value is 5 but only two bits are left in the substituted binary bit stream for expansion. In that case, though it can't be a major issue in time of expansion but it will leads to major problem at the time of decryption as we will get some redundant values. So, we need to keep that redundant information to get the actual message. We keep that information as redundant bit information at the end of expanded binary bit stream. In case of 8-bit sub-key we can keep that information by adding extra eight bits at the end.

Now, by 8-bit (2^8)we can represent a number between 0 to 255,so we can use extra 8-bit to keep redundant bit information up to 256bitssub-key as the key value lies between 0 to255. But we can't keep the information of the redundant bits if the sub-key size is greater than 256.So, solve that problem we will use8-bit if the sub-key is less than or equal to2^8,if sub-key is greater than 2^8then we will use 16-bit to keep that information until the length of sub-keyisgreater2^{16}andsoon.Thatiswewilluse8-bitorthe multipleof8-bitto keep the redundant bit information and this depends on the length of sub-key. So, the redundant bit information will determine by 2^{8n},(here n is1,2,3and soon) where, sub-key>$2^{8(n-1)}$ and sub-key<=2^{8n}.

As the binary expansion of substituted binary bit stream completes and redundant bit information adds we will get the actual expanded binary bit stream. The next technique, method of shuffling will start after that and it take expanded binary bit stream as input. This method has two scenarios. One what type of shuffling case it going to perform? And how many rounds this shuffling will occur? We need10-bitshuffling key to determine that as in our proposed model were strict shuffling cases in four types and shuffling rounds by 255.In this method, the four cases will occur and these are as follows: case1.Unchanged, case2.Swap,case3.Inverse and case4.Swap-Inverse.

*Case1.Unchanged:*In this case we will take expanded binary bit stream as final cipher binary bit stream and convert it to its corresponding cipher text.

Case2.*Swap:*In this case we will take expanded binary bit stream as initial cipher binary bit stream and then perform bit swapping on the basis of key value. Here, key value of our sub-key was 1,3,7,6,5,2,4 and 0.

Let the expanded bit length is24 and we perform 3 shuffling rounds. At the beginning we take the first key

valuewhichis1.Now,in the 1st round we divide the expanded binary bit stream into twoparts,thelefthalfwhichcontains1-bit(as our key valueis1)and the right half which contain rest 23 bits and then swap them. So, it forms a new intermediate cipher binary bit stream and we'll take it as input for next round.

Then, in the 2nd round we take the next key value which is

binary bit stream.

Case3.Inverse: This case is similar as previous. Here we inverse the expanded binary bit stream. If we take the previous conditions then in the 1st round we divide the expanded binary bit stream into two parts, the left half For the 2nd round we divide it into the left half which contains 3-bit (as our key value is 3) and the right half which contain 21b its and inverse those 21 bits. And at the end we divide it into the left half which contains 7-bit (as our key value is 7) and the right half which contain17 bits and inverse those 17bits.After 3rd round we'll get our final cipher binary bit stream.

Case4.Swap-Inverse: This case is nothing but combination of previous methods i.e. Case2 and Case3. Here in each round we perform first the swap and then inverse operation to form the intermediate cipher binary bit stream.

The algorithm for encryption and decryption process in presented next.

Algorithm1 Encryption

Input: input file to be encrypted

K=key

Output: encrypted file

Begin

Step1.Divide the key into three parts such as substitution key, shuffling case and shuffling rounds.

Step2.Take the all character of plain text of the file into a string and generates corresponding binary bit steam on the basis of the size of substitution key

Step3.Substitute input bit stream using substitution key and calculate the key value of that substitution key.

3, again divide it into two parts, the lefthalfwhichcontains3-bit(as our key value is 3) and the right half which contains rest 21 bits and then swap them. And in the last round our key value will be 7 and we'll divide the left half which contains7-bit(as our key value is 7 now) and the right half which contains 17 bits and swap them. So, after 3rd round we'll get our final cipher

which contains1-bit(as our key value is 1)and the right half which containrest23bitsas previous. Then we just inverse the rest of 23 bits and form intermediate cipher binary bit stream.

Step4.Expand substituted binary bit stream on the basis of key value and add redundant bit information in the end of that bit stream.

Step5. Shuffle expanded binary bit stream using shuffling case, shuffling round sand key value information and generates final cipher bit stream.

Go to Step5 until expanded bit stream found null.

Step6. Convert the final cipher bit stream to corresponding cipher text and write it into the encrypted file.

End

Algorithm2 Decryption

Input: input file to be decrypted

K=key

Output: decrypted file

Begin

Step1.Dividethekey into three parts such as substitution key, shuffling case and shuffling rounds, also determine the key value of substitution key.

Step2.Take the all character of cipher text of the file into a string and generates corresponding binary bit steam on the basis of the size of substitution key.

Step3.Shuffle initial cipher bit stream using shuffling case, shuffling rounds and key value information and

generates expanded plain text bit stream.

Go to Step3 until cipher bit stream found null.

Step4. Keep the information of redundant bits into an array and compress expanded plain text bit stream into initial plain text bit stream

Step5. Reconstitute initial plain text bit stream using substitution key and generates corresponding original plain text binary bit stream.

Step6. Convert original plain text binary bit stream to corresponding cipher text and write it into the decrypted file.

End

The length of expanded bit stream is depending of the sub-key, therefore this expanded binary bit stream is varies with the sub-key and it also possible that two different sub-key producing same size of expanded binary bit stream. This approach confuses the unauthorized users about the size of the key. Another important technique we inherit in our model is shuffling cases and shuffling rounds. Here, as we stated, we use four different cases and up to255 rounds which will determine only on the basis of key. Now the major advantage of this mechanism is we can produce various different cipher texts from same expanded binary bit stream by varying different the cases or the rounds.

4. CONCLUSIONS

The paper has taken review of the big data security issues along with the basic properties of Big Data. It shows successful achievement of the encryption and decryption of the given text files. The later part of the paper has presented the encryption and decryption algorithms to demonstrate the security issues of Big Data. These encryption techniques make data more secure. It was observed that size of text file varies to count the effect of volume due to big data.

5. REFERENCES

[1] S.Sen, C.Shaw, R.Chowdhuri, N. Ganguly, and P. Chaudhuri, “Cellular Automata Based Cryptosystem(CAC)”, Fourth International Conference on Information and Communication Security (ICICS02), Dec. 2002, PP. 303-314.

[2] Feng. Bao, ”Cryptanalysis of a Partially Known Cellular Automata Cryptosystem”, VOL.53, No.11, Nov. 2004.

[3] Biham, Eli and Adi Shamir, Differential Cryptanalysis of the Data Encryption Standard, Springer Verlag, 1993.

[4] Coppersmith, D. "The Data Encryption Standard(DES) and Its Strength Against Attacks." IBM Journal of Research and Development, May 1994,pp. 243 - 250.

[5] K. Naik, D. S.L. Wei, Software Implementation Strategies for Power-Conscious Systems,” Mobile Networks and Applications -6, 291-305, 2001.

[6] Farina, A., "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique", presented at the AES 108th Convention, Paris, France, 2002 February 19 -22.

[7] N. Tippenhauer, C. Pöpper, K. Rasmussen, S. Capkun, “On the requirements for successful GPS spoofing attacks,” in Proceedings of the 18th ACM conference on Computer and communications security, pp. 75-86, Chicago, IL, 2011.

[8] P. Gilbert, J. Jung, K. Lee, H. Qin, D. Sharkey, A. Sheth, L. P. Cox, “YouProve: Authenticity and Fidelity in Mobile Sensing,” ACM SenSys 2011, Seattle, WA, November, 2011.

[9] B. Levine, C. Shields, N. Margolin, “A Survey of Solutions to the Sybil Attack,” Tech report 2006-052, University of Massachusetts Amherst, Amherst, MA, October 2006.

[10] B. Agreiter, M. Hafner, and R. Breu, “A Fair Non-repudiation Service in a Web Service Peer-to-Peer

Environment," Computer Standards & Interfaces, vol 30, no 6, pp. 372-378, August 2008.

[11] A. Anagnostopoulos, M. T. Goodrich, and R. Tamassia, "Persistent Authenticated Dictionaries and Their Applications," in Proceedings of the 4th International Conference on Information Security, pp. 379-393, October, 2001.

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 165 - 171, 2016, ISSN:- 2319–8656

Computer Interface for Electroluminescence (EL)

Ajay Kumar Mishra
Department. of Mathematics and Computer Science,
R.D. University,
Jabalpur, India

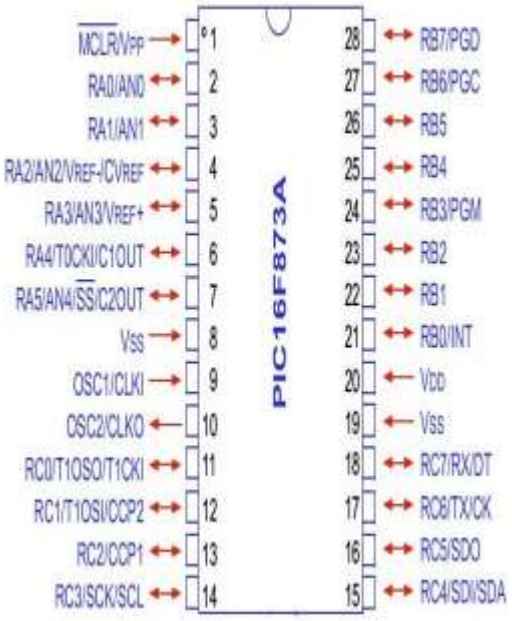
Abstract: The goal of Computer aided device start from the physical description of integrated circuit devices, considering both the physical configuration and related device properties and build the link between the broad range of physics and electrical behavior models that support circuit design. Physics-based modeling of devices, is distributed and lumped form is an essential part of the IC process development. It seeks to quantify the underlying understanding of the technology and abstract that knowledge to the device design level, including extraction of the key parameters that support circuit design and statistical metrology [1][2]. IC development for more than a quarter-century has been dominated by the MOS technology. In the 1970s and 1980s NMOS was favored owing to speed and area advantages, coupled with technology limitations and concerns related to isolation, parasitic effects and process complexity. During that era of NMOS-dominated LSI and the emergence of VLSI, the fundamental scaling laws of MOS technology were codified and broadly applied [3]. It was also during this period that Computer Aided Device reached maturity in terms of realizing robust process modeling (primarily one-dimensional) which then became an integral technology design tool, used universally across the industry [4]. At the same time device simulation, dominantly two-dimensional owing to the nature of MOS devices, became the work-horse of technologists in the design and scaling of devices [5]. The transition from NMOS to CMOS technology resulted in the necessity of tightly coupled and fully 2D simulators for process and device simulations [6][7].

Keywords: **Computer interface, interfacing, computer aided device.**

1. INTRODUCTION

In computer science, an interface is the point of interaction with software, or computer hardware, or with peripheral devices such as a computer monitor or a keyboard. Some computer interfaces such as a touch screen can send and receive data, while others such as a mouse or microphone can only send data [8]. A hardware interfaces exist in computing systems between many of the components such as the various buses, storage devices, other I/O devices, etc. A hardware interface is described by the mechanical, electrical and logical signals at the interface and the protocol for sequencing them (sometimes called signaling). A standard interface, such as SCSI decouples the design and introduction of computing hardware, such as I/O devices, from the design and introduction of other components of a computing system, thereby allowing users and manufacturer's greater flexibility in the implementation of computing systems. Hardware interfaces can be parallel where performance is important or serial where distance is important [9]. A computer network, also referred to as just a network consists of two or more computers and typically other devices as well (such as printers, external hard drives, modems and routers), that are linked together so that they can communicate with each other and thereby exchange commands and share data, hardware and other resources. The devices on a network are referred to as nodes. They are analogous to the knots in nets that have traditionally been used by fishermen and others. Nodes can be connected using any of various types of media, including twisted pair copper wire cable, optical fiber cable, coaxial cable and radio waves. And they can be arranged according to several basic topologies (i.e., layouts), including bus (in which all nodes are connected along a single cable), star (all nodes are connected to a central node), tree (nodes successively branch off from other nodes) and ring

2. DESCRIPTION OF HARDWARE USED



Pin Description of PIC16F873A microcontroller

Pin No.	Pin Name	Description
1	MCLR/VPP	Master Clear (input) or programming voltage

		(output). This pin is active low
2	RAO/ANO	Digital I/O. Analog input 0.
3	RA1/AN1	Digital I/O. Analog input 1.
4	RA2/AN2/VREF-/CVREF	Digital I/O. Analog input 2. A/D reference voltage (Low) input. Comparator VREF output.
5	RA3/AN3/VREF+	Digital I/O. Analog input 3. A/D reference voltage (High) input.
6	RA4/TOCKI/C1OUT	Digital I/O- Open –Drain when configured as output.
7	RA5/AN4/SS/C2OUT	Digital I/O. Analog Input 4. SPI Slave selects input. Comparator 2 output.
8	VSS	Ground reference for logic and I/O Pins.
9	OSC1/CLK1	Oscillator Crystal or external clock input. Oscillator crystal input or external clock source input
10	OSC2/CLKO	Oscillator Crystal or clock output. Oscillator crystal output. Connects to crystal or resonator in crystal oscillator mode.
11	RCO/T1OSO/T1CKI	Digital I/O. Timer1 oscillator output. Timer1 external clock input.
12	RC1/T1OSI/CCP2	Digital I/O. Timer1 oscillator input. Capture2 input, compare2 output, PWM2 output.
13	RC2/CCP1	Digital I/O. Capture1 input, compare 1 output, PWM 1 output.
14	RC3/SCK/SCL	Digital I/O. Synchronous serial Clock input/output for SPI mode.
		Synchronous serial Clock input/output for I^2C mode.
15	RC4/SDI/SDA	Digital I/O. SPI data in. I^2C data I/O.
16	RC5/SDO	Digital I/O. SPI data out.
17	RC6/TX/CK	Digital I/O. USART asynchronous transmit. USART1 synchronous clock.
18	RC7/RX/DT	Digital I/O. USART asynchronous receive. USART synchronous data.
19	VSS	Ground reference for logic and I/O Pins.
20	VDD	Positive Supply for logic and I/O Pins.
21	RBO/INT	Digital I/O. External Interrupts
22	RB1	Digital I/O.
23	RB2	Digital I/O.
24	RB3/PGM	Digital I/O. Low- voltage (single-supply) ICSP programming enable pin.
25	RB4	Digital I/O.
26	RB5	Digital I/O.
27	RB6/PGC	Digital I/O. In-circuit debugger and ICSP programming clock.
28	RB7/PGD	Digital I/O. In-circuit debugger and ICSP programming data.

MICROCONTROLLER PROGRAMING

PIC16F873A microcontroller used program code as given below

……
……

```
LIST P=PIC16F873A
        INCLUDE "P16F873A.INC"
BANK0 EQU 20H
```

```
        CBLOCK BANK0
        LOWERLSB
        HIGHER
        UNIT
        TEN
        HUND
        R1
        R2
        REQUEST
        HEX
        THOU
        TEMP
        DIGIT_SEL
        DIGIT_DISP
        DIGIT_OUT
        REPEAT
        ENDC
        ORG 0X000
RVRESET
        GOTO START
        ORG 0X004
RVINT
        BTFSS PIR1, ADIF
        GOTO $-1
        BCF PIR1, ADIF
        CALL ADC_INT
        GOTO PROGRAM
        RETFIE

START
        CALL SET_PORTS
        MOVLW 0X0F
        MOVWF REPEAT
        GOTO PROGRAM

SET_PORTS
        ;CALL SET_PORTA
        CALL SET_PORTB
        CALL SET_PORTC
        RETURN
SET_PORTA
        BANKSEL ADCON1
        MOVLW 0X06
        MOVWF ADCON1
        BANKSEL TRISA
        MOVLW 0X3F
        MOVWF TRISA
        BANKSEL PORTA
        CLRF PORTA
        RETURN
SET_PORTB
        BANKSEL TRISB
        CLRF TRISB
        BANKSEL PORTB
        CLRF PORTB
        RETURN
SET_PORTC
        BANKSEL TRISC
        CLRF TRISC
        BANKSEL PORTC
        CLRF PORTC
        RETURN
PROGRAM
        CALL SET_PORTB
        CALL SET_PORTC
        CALL HEX_TO_BCD
        CALL NUMBER
        CALL DISPLAY
        DECFSZ REPEAT, F
```

```
        GOTO $-2
        MOVLW 0X0F
        MOVWF REPEAT
        CALL REQUEST_1
        CALL ADC_1
        GOTO PROGRAM
REQUEST_1
        MOVLW 0X01
        MOVWF REQUEST
        RETURN
REQUEST_2
        MOVLW 0X09
        MOVWF REQUEST
        RETURN
REQUEST_3
        MOVLW 0X11
        MOVWF REQUEST
        RETURN
REQUEST_4
        MOVLW 0X19
        MOVWF REQUEST
        RETURN
ADC_1
        MOVLW 0XC0
        MOVWF INTCON
        MOVF REQUEST, W
        MOVWF ADCON0
        CLRF PORTA
        CLRF ADRESH
        BANKSEL PIE1
        CLRF PIE1
        BSF PIE1, ADIE
        MOVLW 0X40
        MOVWF ADCON1
        CLRF TRISA
        COMF TRISA, F
        CLRF ADRESL
        NOP
        NOP
        NOP
        NOP
        BANKSEL ADCON0
        MOVLW 0X05
        MOVWF ADCON0
        GOTO $

ADC_INT
        CALL LOW_1
        CALL HIGH_1
        RETURN
LOW_1
        BANKSEL ADRESL
        MOVF ADRESL, W
        MOVWF LOWERLSB
        RETURN
HIGH_1
        BANKSEL ADRESH
        MOVF ADRESH, W
        MOVWF HIGHER
        MOVWF HEX
        RETURN
OUTPUT
        BANKSEL TRISB
        CLRF TRISB
        BANKSEL PORTB
        MOVF HIGHER, W
        MOVWF PORTB
        RETURN
```

```
NUMBER
        SWAPF UNIT, W
        ANDLW 0X0F
        MOVWF TEN
        MOVLW 0X0F
        ANDWF UNIT, F
        RETURN
HEX_TO_BCD
        BANKSEL STATUS
        BCF STATUS, 0    ;1 shift
        RLF HEX, F
        RLF UNIT, F
        BCF STATUS, 0          ;2 shift
        RLF HEX, F
        RLF UNIT, F
        BCF STATUS, 0          ;3 shift
        RLF HEX, F
        RLF UNIT, F
        CALL UNIT_5
        BCF STATUS, 0          ;4 shift
        RLF HEX, F
        RLF UNIT, F
        CALL UNIT_5
        BCF STATUS, 0          ;5 shift
        RLF HEX, F
        RLF UNIT, F
        CALL UNIT_5
        BCF STATUS, 0          ;6 shift
        RLF HEX, F
        RLF UNIT, F
        CALL UNIT_5
        CALL TEN_5

        BCF STATUS, 0          ;7 shift
        RLF HEX, F
        RLF UNIT, F
        RLF HUND, F
        CALL UNIT_5
        CALL TEN_5
        BCF STATUS, 0          ;8 shift
        RLF HEX, F
        RLF UNIT, F
        RLF HUND, F
        RETURN
UNIT_5
        MOVF UNIT, W
        ANDLW 0X0F
        MOVWF TEMP
        MOVLW 0X05
        SUBWF TEMP, W
        BANKSEL STATUS
        BTFSS STATUS, 0
        RETURN
        MOVLW 0X03
        ADDWF UNIT, F
        RETURN
TEN_5
        MOVF UNIT, W
        ANDLW 0XF0
        MOVWF TEMP
        MOVLW 0X50
        SUBWF TEMP, W
        BANKSEL STATUS
        BTFSS STATUS, 0
        RETURN
        MOVLW 0X30
        ADDWF UNIT, F
        RETURN
```

```
DISPLAY
        CLRF DIGIT_SEL
        BCF DIGIT_SEL, 7
        BSF DIGIT_SEL, 4
        MOVF UNIT, W
        MOVWF DIGIT_DISP
        CALL S_S_DECODER
        MOVWF DIGIT_OUT
        CALL DIGIT_SELECT
        CALL OUT_TO_FND
        CALL DELAY
        BCF DIGIT_SEL, 4
        BSF DIGIT_SEL, 5
        MOVF TEN, W
        MOVWF DIGIT_DISP
        CALL S_S_DECODER
        MOVWF DIGIT_OUT
        CALL DIGIT_SELECT
        CALL OUT_TO_FND
        CALL DELAY

        BCF DIGIT_SEL, 5
        BSF DIGIT_SEL, 6
        MOVF HUND, W
        MOVWF DIGIT_DISP
        CALL S_S_DECODER
        MOVWF DIGIT_OUT
        CALL DIGIT_SELECT
        CALL OUT_TO_FND
        CALL DELAY
        BCF DIGIT_SEL, 6
        BSF DIGIT_SEL, 7
        MOVF THOU, W
        MOVWF DIGIT_DISP
        CALL S_S_DECODER
        MOVWF DIGIT_OUT
        CALL DIGIT_SELECT
        CALL OUT_TO_FND
        CALL DELAY
        GOTO DISPLAY
        RETURN
S_S_DECODER
        ;Display code table.....
        MOVF DIGIT_DISP, W ; Get key count
        ADDWF PCL ; and calculate jump
        ;NOP ; into table
        RETLW B'11011110' ; Code for '0'
        RETLW B'01000010' ; Code for '1'
        RETLW B'11101100' ; Code for '2'
        RETLW B'11100110' ; Code for '3'
        RETLW B'01110010' ; Code for '4'
        RETLW B'10110110' ; Code for '5'
        RETLW B'10111110' ; Code for '6'
        RETLW B'11000010' ; Code for '7'
        RETLW B'11111110' ; Code for '8'
        RETLW B'11110110' ; Code for '9'
        ;Output display code.....
        RETURN
OUT_TO_FND
        MOVF DIGIT_OUT, W
        MOVWF PORTB
        RETURN
DIGIT_SELECT
        CLRF PORTB
        MOVF DIGIT_SEL, W
        MOVWF PORTC
        RETURN
DELAY
```

```
        MOVLW 0X0F
        MOVWF R1
        MOVWF R2
        DECFSZ R2, F
        GOTO $-1
        DECFSZ R1, F
        GOTO $-4
        RETURN
DELAY_L
        MOVLW 0XFF
        MOVWF R1
        MOVWF R2
        DECFSZ R2,F
        GOTO $-1
        DECFSZ R1,F
        GOTO $-4
        RETURN
            END
```

3. MEASUREMENT OF ELECTROLUMINESCENCE (EL)

For the measurement of EL brightness EL cell is prepared by selected method and it is ready to use for measurement of luminescence. In prepared cell, there are two electrodes, one from conducting glass plate and other from phosphor sample side. Frequency is applied to amplifier at the desired level and it increases voltage up to required level, when frequency and voltage are reached at certain level, then emission of EL is takes place the emission of light is connected to Photomultiplier tube (PMT) and it converts the light in the form of current and then converted to voltage by voltage to current converter. The output voltage is apply to PIC16F873A microcontroller and it convert this analog voltage to Digital voltage. The digital voltage to be interfaced with the computer by parallel port. The data of parallel port is read by VB.NET software and finally luminescence of EL cell is converted in voltage, and is obtained at the screen of the computer.

Figure 1 Sample of Interfacing Window

4. REFERENCES

[1] H.J. DeMan and R. Mertens, SITCAP--A simulator of bipolar transistors for computer-aided circuit analysis programs, International Solid-State Circuits Conference (ISSCC), Technical Digest, pp. 104-5, February, 1973

[2] R.W. Dutton and D.A. Antoniadis, Process simulation for device design and control, International Solid-State Circuits Conference (ISSCC), Technical Digest, pp. 244-245, February, 1979

[3] R.H. Dennard, F.H. Gaensslen, H.N. Yu, V.L. Rodeout, E. Bassous and A.R. LeBlanc, Design of ion-implanted MOSFETs with very small physical dimensions, IEEE Jour. Solid-State Circuits, vol. SC-9, pp.256-268, October, 1974.

[4] R.W. Dutton and S.E. Hansen, Process modeling of integrated circuit device technology, Proceeding IEEE, vol. 69, no. 10, pp. 1305-1320, October, 1981.

[5] P.E. Cottrell and E.M. Buturla, "Two-dimensional static and transient simulation of mobile carrier transport in a semiconductor," Proceedings NASECODE I (Numerical Analysis of Semiconductor Devices), pp. 31-64, Boole Press, 1979.

[6] C.S. Rafferty, M.R. Pinto, and R.W. Dutton, Iterative methods in semiconductor device simulation, IEEE Trans. Elec. Dev., vol. ED-32, no.10, pp.2018-2027, October, 1985.

[7] M.R. Pinto and R.W. Dutton, Accurate trigger condition analysis for CMOS latchup, IEEE Electron Device Letters, vol. EDL-6, no. 2, February, 1985.

[8] IEEE 100 - The Authoritative Dictionary Of IEEE Standards Terms. NYC, NY, USA: IEEE Press. 2000. pp. 574–575. ISBN 0-7381-2601-2.

[9] Blaauw, Gerritt A.; Brooks, Jr., Frederick P. (1997), "Chapter 8.6, Device Interfaces", Computer Architecture-Concepts and Evolution, Addison-Wesley, pp. 489–493, ISBN 0-201-10557-8 See also:

Patterson, David A.; Hennessey, John L. (2005), "Chapter 8.5, Interfacing I/O Devices to the Processor, Memory and Operating System", Computer Organization and Design - The Hardware/Software Interface, Third Edition, Morgan Kaufmann, pp. 588–596, ISBN 1-55860-604-1

International Journal of Computer Applications Technology and Research
Volume 5– Issue 3, 172 - 175, 2016, ISSN:- 2319–8656

A Comparative Study of Various Data Mining Techniques: Statistics, Decision Trees and Neural Networks

Balar Khalid
Department of Computer Science
Hassan II University-FMPC
Casablanca, Morocco

Naji Abdelwahab
Department of Computer Science
Hassan II University-ENSET
Mohammedia, Morocco

Abstract: In this paper we focus on some techniques for solving data mining tasks such as: Statistics, Decision Trees and Neural Networks. The new approach has succeed in defining some new criteria for the evaluation process, and it has obtained valuable results based on what the technique is, the environment of using each techniques, the advantages and disadvantages of each technique, the consequences of choosing any of these techniques to extract hidden predictive information from large databases, and the methods of implementation of each technique. Finally, the paper has presented some valuable recommendations in this field.

Keywords: Data mining, Statistics, Logistic Regression, Decision Trees and Neural Networks.

1. INTRODUCTION

Extraction useful information from data is very far easier from collecting them. Therefore many sophisticated techniques, such as those developed in the multi- disciplinary field data mining are applied to the analysis of the datasets. One of the most difficult tasks in data mining is determining which of the multitude of available data mining technique is best suited to a given problem. Clearly, a more generalized approach to information extraction would improve the accuracy and cost effectiveness of using data mining techniques.

Therefore, this paper proposes a new direction based on evaluation techniques for solving data mining tasks, by using three techniques: Statistics, Decision Tree and Neural Networks.

The aim of this new approach is to study those techniques and their processes and to evaluate data mining techniques on the basis of: the suitability to a given problem, the advantages and disadvantages, and the consequences of choosing any technique, [5].

2. DATA MINING TOOLS

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [6]. Data mining tools predict future trends and behaviors allowing businesses to make proactive knowledge driven decisions. Data mining tools can answer business question that traditionally were too time consuming to resolve.

They scour database for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

3. SELECTED DATA MINING TECHNIQUES

A large number of modeling techniques are labeled "data mining" techniques [7]. This section provides a short review of a selected number of these techniques. Our choice was guided the focus on the most currently used models. The review in this section only highlights some of the features of different techniques and how they influence, and benefit from. We do not present a complete exposition of the mathematical details of the algorithms, or their implementations.

Although various different techniques are used for different purposes those that are of interest in the present context [4]. Data mining techniques which are selected are Statistics, Decision Tree and Neural Networks.

3.1 Statistical Techniques

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models.

Today people have to deal with up to terabytes of data and have to make sense of it and glean the important patterns from it. Statistics can help greatly in this process by helping to answer several important questions about their data: what patterns are there in database?, what is the chance that an event will occur?, which patterns are significant?, and what is a high level summary of the data that gives some idea of what is contained in database?

In statistics, prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction.

The simplest form of regression is S*imple Linear Regression* that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line [2].

Adding more predictors to the linear equation can produce more complicated lines that take more information into

account and hence make a better prediction, and it is called multiple linear regressions.

3.2 Decision Tree Techniques

The decision tree is a predictive model that, as its name implies, can be viewed as a decision tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object. Induction decision tree can be used for exploration analysis, data preprocessing and prediction work.

The process in induction decision tree algorithms is very similar when they build trees. These algorithms look at all possible distinguishing questions that could possibly break up the original training dataset into segments that are nearly homogeneous with respect to the different classes being predicted. Some decision tree algorithms may use heuristics in order to pick the questions. As example, CART (Classification And Regression Trees) picks the questions in a much unsophisticated way as it tries them all. After it has tried them all, CART picks the best one, uses it to split the data into two more organized segment and then again ask all possible questions on each of these new segment individually [4].

3.3 Neural Network Technique

Artificial neural network derive their name from their historical development which started off with the premise that machines could be made to think if scientists found ways to mimic the structure and functioning of the human brain on the computer. There are two main structures of consequence in the neural network: The node - which loosely corresponds to the neuron in the human brain and the link - which loosely corresponds to the connections between neurons in the human brain [4].

Therefore, a neural network model is a collection of interconnected neurons. Such interconnections could form a single layer or multiple layers. Furthermore, the interconnections could be unidirectional or bi-directional. The arrangement of neurons and their interconnections is called the architecture of the network. Different neural network models correspond to different architectures. Different neural network architectures use different learning procedures for finding the strengths of interconnections.

Therefore, there are a large number of neural network models; each model has its own strengths and weaknesses as well as a class of problems for which it is most suitable.

4. EVALUATION OF DATA MINING TECHNIQUES

In this section, we can compare the selected techniques with the five criteria [5]: The identification of technique, the environment of using each technique, the advantages of each technique, the disadvantages of each technique, the consequences of choosing of each technique, and the implementation of each technique's process.

4.1 Statistical Technique

4.1.1 Identification of Statistics

"Statistics is a branch of mathematics concerning the collection and the description of data" [2].

4.1.2 The Environment of Using Statistical Technique

Today data mining has been defined independently of statistics though "mining data" for patterns and predictions is really what statistics is all about. Some of the techniques that are classified under data mining such as CHAID and CART really grew out of the statistical profession more than anywhere else, and the basic ideas of probability, independence and causality and over fitting are the foundation on which both data mining and statistics are built. The techniques are used in the same places for the same types of problems (prediction, classification discovery).

4.1.3 The Advantages of Statistical Technique

Statistics can help greatly in data mining process by helping to answer several important questions about your data. The great values of statistics is in presenting a high level view of the database that provides some useful information without requiring every record to be understood in detail. As example, the histogram can quickly show important information about the database, which is the most frequent.

4.1.4 The Disadvantages of Statistical Technique

Certainly statistics can do more than answer questions about the data but for most people today these are the questions that statistics cannot help answer. Consider that a large part of data the statistics is concerned with summarizing data, and more often than not, the problem that the summarization has to do with counting.

Statistical Techniques cannot be useful without certain assumptions about data.

4.1.5 The Consequences of choosing The Statistical Technique

Statistics is used in the reporting of important information from which people may be able to make useful decisions. A trivial result that is obtained by an extremely simple method is called a naïve prediction, and an algorithm that claims to learn anything must always do better than the naïve prediction.

4.2 Decision Trees Technique

4.2.1 Identification of Decision Trees

"A decision tree is a predictive model that, as its name implies, can be viewed as a tree " [2].

4.2.2 The Environment of using Decision Trees Technique

Decision trees are used for both classification and estimation tasks. Decision trees can be used in order to predict the outcome for new samples. The decision tree technology can be used for exploration of the dataset and business problem. Another way that the decision tree technology has been used is for preprocessing data for other prediction algorithms.

4.2.3 The Advantages of Decision Trees Technique

The Decision trees can naturally handle all types of variables, even with missing values. Co-linearity and linear-separability problems do not affect decision trees performance. The representation of the data in decision trees form gives the illusion of understanding the causes of the observed behavior of the dependent variable.

4.2.4 The Disadvantages of Decision Trees Technique

Decision trees are not enjoying the large number of diagnostic tests. Decision trees do not impose special restrictions or requirements on the data preparation procedures. Decision trees cannot match the performance of that of linear regression.

4.2.5 Consequences of choosing of Decision Trees Technique

The decision trees help to explain how the model determined the estimated probability (in the case of classification) or the mean value (in the case of estimation problems) of the dependent variable. Decision trees are fairly robust with respect to a variety of predictor types and it can be run relatively quickly. Decision trees can be used on the first pass of a data mining run to create a subset of possibly useful predictors that can then be fed into neural networks, nearest neighbor and normal statistical routines.

4.3 Neural Networks Technique

4.3.1 Identification of Neural Network

"A neural network is given a set of inputs and is used to predict one or more outputs". [3]. "Neural networks are powerful mathematical models suitable for almost all data mining tasks, with special emphasis on classification and estimation problems" [9].

4.3.2 The Environment of using Neural Networks Technique

Neural network can be used for clustering, outlier analysis, feature extraction and prediction work. Neural Networks can be used in complex classification situations.

4.3.3 The Advantages of Neural Networks Technique

Neural Networks is capable of producing an arbitrarily complex relationship between inputs and outputs.
Neural Networks should be able to analyze and organize data using its intrinsic features without any external guidance. Neural Networks of various kinds can be used for clustering and prototype creation.

4.3.4 The Disadvantages of Neural Networks Technique

Neural networks do not work well when there are many hundreds or thousands of input features. Neural Networks do not yield acceptable performance for complex problems. It is difficult to understand the model that neural networks have built and how the raw data affects the output predictive answer.

4.3.5 Consequences of choosing of Neural Networks Technique

Neural Networks can be unleashed on your data straight out of the box without having to rearrange or modify the data very much to begin with. Neural Networks is that they are automated to a degree where the user does not need to know that much about how they work, or predictive modeling or even the database in order to use them.

5. CONCLUSION

In this paper we described the processes of selected techniques from the data mining point of view. It has been realized that all data mining techniques accomplish their goals perfectly, but each technique has its own characteristics and specifications that demonstrate their accuracy, proficiency and preference.
We claimed that new research solutions are needed for the problem of categorical data mining techniques, and presenting our ideas for future work.
Data mining has proven itself as a valuable tool in many areas, however, current data mining techniques are often far better suited to some problem areas than to others, therefore it is recommend to use data mining in most companies for at least to help managers to make correct decisions according to the information provided by data mining.
There is no one technique that can be completely effective for data mining in consideration to accuracy, prediction, classification, application, limitations, segmentation, summarization, dependency and detection. It is therefore recommended that these techniques should be used in cooperation with each other.

6. REFERENCES

[1] Adamo, J. M, Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms Springer-Verlag, New York, 2001.

[2] Berson, A, Smith, S, and Thearling, K. Building Data Mining Applications for CRM, 1st edition - McGraw-Hill Professiona, 1999.

[3] Bramer, M. Principles of Data Mining, Springer-Limited, 2007.

[4] Dwivedi, R. and Bajpai, R. Data Mining Techniques for dynamically Classifying and Analyzing Library Database Convention on Automation of Libraries in Education and Research Institutions, CALIBER, 2007.

[5] El-taher, M. Evaluation of Data Mining Techniques, M.Sc thesis (partial-fulfillment), University of Khartoum, Sudan, 2009.

[6] Han, J and Kamber, M. Data Mining , Concepts and Techniques, Morgan Kaufmann , Second Edition, 2006.

[7] Lee, S and Siau, K. A review of data mining techniques, Journal of Industrial Management & Data Systems, vol 101, no 1, 2001, pp.41-46.

[8] Perner, P. Data Mining on Multimedia - Springer- Limited , 2002.

[9] Refaat, M. Data Preparation for Data Mining Using SAS, Elsevier, 2007.

[10] Vityaev, E and Kovalerchuk, B. Inverse Visualization In Data Mining, in International Conference on Imaging Science, Systems, and Technology CISST'02, 2002.

BALAR Khalid[1], PhD in Computer Science, Hassan II University-Faculty of Medicine and Pharmacy of Casablanca, 19 Rue Tarik Ibnou Ziad, B.P. 9154, Casablanca, Morocco. Email: balarkhalid@gmail.com

NAJI Abdelwahab[2]**,** Assistant Prof in Computer Science, Hassan II University- Superior Normal School of Technical Education, Rue Boulevard Hassan II Mohammedia, Morocco. Email: abdelwahab.naji@gmail.com