Journal of Advances in Information Technology

ISSN 1798-2340

Volume 3, Number 4, November 2012

Contents

Comparison of Segmentation Tools for Multiple Modalities in Medical Imaging

Sonali Bhadoria
MAE, Pune, India
Sonali_nakade@yahoo.com

Preeti Aggarwal
UIET, Panjab University, Chandigarh, India
pree_agg2002@yahoo.com

C. G. Dethe
PIET, Nagpur, India
cgdethe@yahoo.com

Renu Vig
UIET, Panjab University, Chandigarh, India
renuvig@hotmail.com

***Abstract*—Image segmentation plays a crucial role in many medical imaging applications by extracting the regions of interest. Accurate segmentation of medical images is a key step in the use of computer-aided diagnosis (CAD) systems to improve the sensitivity and specificity of lesion detection. In this paper, segmentation problems in medical imaging modalities especially for lung CT as well as for thyroid ultrasound (US) are discussed along with their comparative results are shown using automatic tools as well as with some specific algorithms. In this paper various automatic tools as well as manual segmentation algorithms have been used and compared. Both the outcomes either from automatic tool as well as using an algorithm provide the required ROI (region of interest) but automatic tool's output is more efficient and perfect. 3D visualization as well as volumetric segmentation is done accurately with the help of these tools which help in segmenting CT (3D) images especially.**

***Index Terms*—CT, US, Region of Interest (ROI), Interstitial Lung Disease (ILD).**

I. INTRODUCTION

Medical Imaging technologies, such as CT, MRI, Positron emission tomography (PET), US have been widely applied to various medical procedures. Compared to traditional medical diagnosis, they provide non-invasive yet powerful means to investigate the internal structures and activities of human bodies. With the help of such technologies [1] , doctors can obtain multi-dimensional information such as 2-D slices, 3-D volumetric images and videos of ROI, which facilitates the performance of both qualitative and quantitative analysis. Segmentation is a main domain of medical image processing [2]. Medical image segmentation is usually the first step of most analysis procedures [3]. It is often important to separate regions or objects of interest from other parts of the body. If we can define the borders of the ROI (segment the image) we can often simplify the decisions [3]. It is also a crucial step that determines the final result of the entire application, since the rest of the analysis fully relies on the output from segmentation step.

Methods for performing segmentations vary widely depending on the specific application, imaging modality, and other factors [4]. For example, the segmentation of brain tissue has different requirements from the segmentation of the liver. General imaging artifacts such as noise, partial volume effects, and motion can also have significant consequences on the performance of segmentation algorithms. Any approaches for the segmentation proposed in literature vary widely depending on the specific application, imaging modality CT, MRI, US etc.), and other factors. The algorithm, which gives perfect results for one application, might not even work for another. Mostly, segmentation is semi-automatic and a seed point is needed. Then, the structure is being segmented as exactly as possible, for example to measure its size, volume or form, in the case of a tumor. Even now a day's researchers as well physician's try to make this crucial step as easy as possible either with the help of automatic software tools or my means of any algorithm. In this paper, after discussing two most important imaging modalities CT and US, we have discussed and compared the outputs of CT lung and US thyroid segmentation using automatic tools as well as specific algorithm. Various automatic tools have been used

doi:10.4304/jait.3.4.197-205

like Mazda [5], YaDiv [6] as well Analyze 10.0 [7]. Also MATLAB based tool MATITK [8] has also been evaluated for segmentation.

II. Medical Image Modalities

Various medical imaging are available like MRI, CT, US, positron emission tomography (PET), etc. depending upon need, disease type and body organ. Here we will discuss the two most common imaging modality CT and US in details.

A. Computed Tomography (CT)

CAD with CT data [9] [10] can increase the radiologist's efficacy and provide more accurate diagnosis for lung cancer. Computed Tomography, also known as computed axial tomography, or CAT scan is a medical technology that uses X-rays and computers to produce three-dimensional images of the human body. Unlike traditional X-rays, which highlight dense body parts, such as bones, CT provides detailed views of the body's soft tissues, including blood vessels, muscle tissue, and organs, such as the lungs. While conventional X-rays provide flat two-dimensional images, CT images depict a cross-section of the body which helps in detecting various lung diseases like ILD as well as tumors. Due to the development of multi-slice CT technology, a modern CT scanner [11] can now generate a large number (500–1000) of slices for each patient's CT image scan, covering a large volume of the human body within a short time. Based on this high performance, radiologists can easily photograph the whole human chest, abdomen, or torso with high spatial resolution in a one-time CT scan. Multisided CT imaging is the primary digital technique for imaging the lung for the detection of pulmonary (lung) disease such as lung cancer, tumor, and cystic fibrosis [12]. Sometimes doctors recommend the MRI of lung depending upon the patient's condition. Figure 1 and Figure 2 shows lung cancer in both MRI as well as in CT scan respectively.

Figure 1. 322x 286 MRI Scan of chest showing cancer

Figure 2. 1280x 1025 CT scan of 72 years old woman with lung cancer

B. Ultrasounds

Medical ultrasound, also called sonography, is a mode of medical imaging that has a wide array of clinical applications, both as a primary modality and as an adjunct to other diagnostic procedures. The basis of its operation is the transmission of high frequency sound into the body followed by the reception, processing, and parametric display of echoes returning from structures and tissues within the body. US are an ideal imaging modality for detection and assessment of a thyroid nodule. It is easy to perform, widely available and does not involve ionizing radiation. The use of high frequency transducers has significantly improved the spatial and contrast resolution in evaluating superficial structures including the thyroid gland.

Ultrasound imaging of thyroid gland provides the ability to acquire valuable information for medical diagnosis. Physicians usually diagnose the pathology of the thyroid gland by its volume. However, even if the thyroid glands are found and the shapes are hand-marked from US images, most physicians still depend on CT images, which are expensive to obtain, for precise measurements of the volume of the thyroid gland and detection of nodules in it. This approach relies heavily on the experience of the physicians and is very time consuming. US imaging is thus one of the most commonly used auxiliary tools in clinical diagnosis. This test is helpful in determining if a thyroid nodule is solid or filled with fluid. It can also be used to check the number and size of thyroid nodules. Ultrasound features can sometimes suggest a nodule is likely to be cancerous, but can't predict malignancy for certain [13]. Figure 3 shows the prese[illegible]id image.

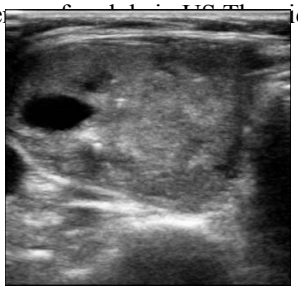


Figure 3. Ultrasound Image of Thyroid Showing Nodule

III. Problems In Segmentation

Segmentation has an important role in medical imaging as it helps in extracting the organ of interest. For the diagnosis of lung it is necessary to segment the chest images and extracts the lungs and further the nodule. Following are the problems generally noticed while segmenting CT lung and US thyroid.

A. CT Lung

Here, we are listing some of the problems faced by physicians as well as the researchers while segmenting CT lung images:

i. *First*, for Ground Glass Opacity (GGO) nodules [14] , the low contrast and fuzzy margins make accurate segmentation of GGO very hard.
ii. *Second*, the usually long duration (one year for example) and the different conditions between two sets of CT scan cause large non-rigid deformation of lung, and intensity differences within the same tissue.
iii. *Thirdly*, the large data size of high-resolution CT scanning can cause problems for 3D segmentation, due to the need of significant memory and computational resources.
iv. Inferior soft tissue contrast compared to MRI as it is X-ray-based.
v. Conventional methods of lung segmentation rely on a large gray value contrast between lung fields and surrounding tissues. These methods fail on scans with lungs that contain dense pathologies, and such scans occur frequently in clinical practice.

Despite of all the above problems, still segmentation of CT images plays an important role in final diagnosis of a disease especially nodule detection and growth.

B. *US Thyroid*

Despite of various advantages in using US images for detection of nodules there are few problems faced during segmentation as follows [15].

i. US imaging suffers from the presence of a granular pattern termed as speckle. Due to this finding accurate texture features for segmentation gets difficult.
ii. There are random fluctuations in the image's intensity profile. Reverberation, shadowing, refraction, side and grating lobes deteriorate the resolution of the US image, thus degrade its overall quality. This cause extraction of spatial and statistical features difficult.
iii. During analyzing the image, muscles present in image may be misinterpreted as a nodule because it gives very similar visual effects in US images.
iv. Boundary of the image is not fixed as it is dependent on the angle of image taken.

Taking in consideration all the above reasons, a correct boundary estimation of a thyroid nodule may play a key role in thyroid US segmentation.

IV. SEGMENTATION METHODS FOR LUNG CTS

For the detection of lung diseases generally X-rays of lungs are performed but for more details physicians recommend the CT scan of the patient's lung. The purpose of the segmentation of the lung region in the CT image is to achieve a better orientation in the image. A lot of articles can be found regarding segmentation of the lung region in CT images. Now, it's necessary to understand the lung structure before discussing how to segment the lung CT images.

Lung Structure

The lung, the site of gas exchange, is filled with air that has a low density (about −1000 HU) on CT images. In addition to air, pulmonary vessels and bronchi are the principal constituents of the lung regions. Lung regions include the left and right lungs. The left lung is further separated into two lung lobes (upper lobe and lower lobe) by an oblique fissure. The right lung is separated into three lung lobes (upper lobe, middle lobe, and lower lobe) by oblique and horizontal fissures. The clinical CT image of the lung is shown in Figure 4 and its general anatomy is shown in Figure 5.

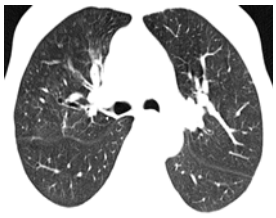


Figure 4. Clinical CT Image

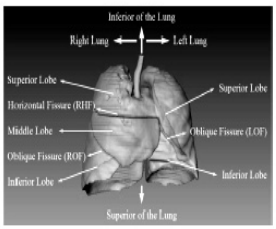


Figure 5. General anatomy of Human lung

A. *Lung Segmentation Using Automatic Tools*

The preprocessing step of most CAD systems for identifying the lung diseases is lung segmentation. The goal of this step is to separate human body regions from background and make an initial classification showing the right and left lung clearly. After that for the detection of any lung disease it is further segmented to extract the exact ROI like nodules in the case of cancer detection. There can be more than one nodule in a lung, so that much number of segmentation steps has to be applied. In this paper, the outputs from various automatic tools like Analyze, Mazda, YaDiv, and MATITK on CT is discussed and compared. The results show that segmentation using these tools makes the segmentation process better, easier as well efficient.

a. Segmentation of Lung using Analyze:

Analyze 10.0 is a powerful, comprehensive software package for multi-dimensional display, processing, and measurement of multi-modality biomedical images. The product of more than 25 years of biomedical imaging research and development at Mayo Clinic, this integrated, total solution allows us to significantly enhance our

multidimensional biomedical imaging productivity [7]. Following figures

Figure 6(a)-6(d) shows the output from Analyze 10.0.

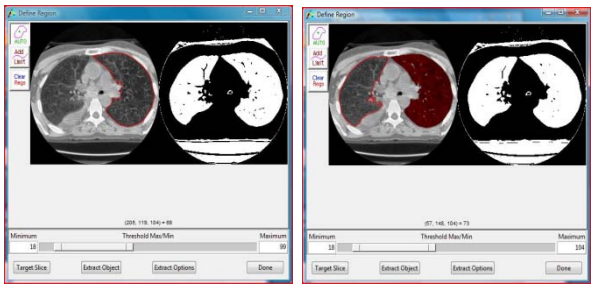


Figure 6 (a): Thresholding using Analyze
Figure 6(b): Region filling using Analyze

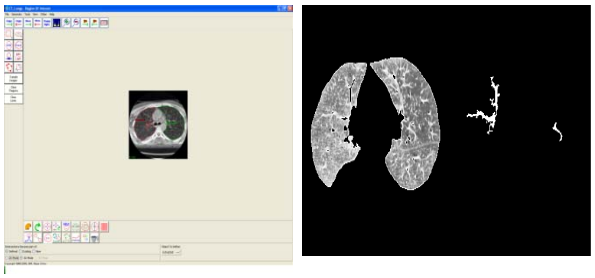


Figure 6(c), (d): Original housefield value and labeling ROI; Nodule Detection using Analyze using Analyze

It can be seen that Analyze is an interactive as well user friendly tool as thresholding, region filling, ROI detection, ROI labeling is really very easy and fast. Figure6 (e) show the various low level features that can be extracted using Analyze. The output of Analyze can be easily exported to other platforms and make Analyze more flexible.

Sat Jun 19 12:33:21 MHT 2010
VolumeFile= CT_Lungs2
SampleMax= 104 SampleMin= 0
VoxelWidth= 1 VoxelHeight= 1 VoxelDepth= 1
#

# Vol_#	Slice	Name	Mean	Std.Dev.	Entropy	Voxels	Area_mm2	Vol_mm3
1	107	nodule1	0.00	0.00	0.00	93	93.00	93.00
1	108	nodule1	22.25	41.49	3.04	93	93.00	93.00
1	109	nodule1	61.13	46.57	4.09	93	93.00	93.00
1	110	nodule1	74.13	31.83	4.37	93	93.00	93.00
1	111	nodule1	71.90	28.10	4.40	93	93.00	93.00
1	112	nodule1	77.51	21.08	4.47	93	93.00	93.00
1	113	nodule1	76.71	20.57	4.48	93	93.00	93.00
1	114	nodule1	74.77	19.31	4.48	93	93.00	93.00
1	115	nodule1	76.66	13.70	4.52	93	93.00	93.00
1	116	nodule1	74.83	12.21	4.52	93	93.00	93.00
1	117	nodule1	73.16	11.68	4.52	93	93.00	93.00

Figure 6 (e): Feature calculation using Analyze

b. Segmentation of lung using MaZda:

MaZda is a computer program for calculation of texture parameters (features) in digitized images. It has been under development since 1998, to satisfy the needs of the participants of COST B11 European project. The program code has been written in C++ and Delphi. The statistical parameters computed by

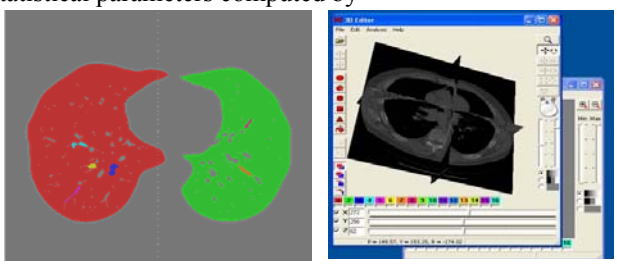


Figure 7 (a): 8 ROI detection in chest CT
Figure 7 (b): 3d view of chest CT

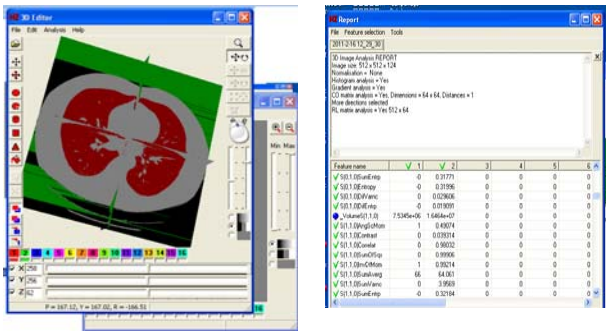


Figure 7 (c): 3d ROI detection of chest CT
Figure 7 (d): List of features calculated by MaZda

the early version of the program were derived from the co-occurrence matrix. Consequently, the name of the program is an acronym derived from 'Macierz Zdarzen' that is the Polish counterpart of the English term 'co-occurrence matrix'. Mazda has proved to be an efficient and well known tool for liver [17], brain [17]. The output from these research shows that it has really helped the physicians to provide the right diagnosis at right time. Figure 7(a)-(e) shows the results of Mazda on chest CT scan. Figure 7(d) shows the few of low level features that can be extracted using MaZda out of 300 features.

c. Segmentation of Lung using Yadiv:

A recent developed DICOM viewer YaDiV [2] has been evaluated for identification of various lung tissues as well as for efficient visualization of lung images. Segmentation of the lung volumes is a required preliminary step to lung tissue categorization. Since the geometries and shapes of the lungs are subject to large variations among the cases, semi–automatic segmentation based on region growing and mathematical morphology is used. The range and region growing routine contained in YaDiv is tested. The resulting binary mask *Mlung* describes the global lung regions well but contains many holes where the region growing algorithm was stopped by denser regions. To fill these holes, a closing operation is

applied to *Mlung* using a spherical structuring element. Then, based on the volumes of the segmented tissues and a set of selected clinical parameters, similar cases are retrieved from a multimedia database of ILD cases as shown in Figure 8 (a) and (b).

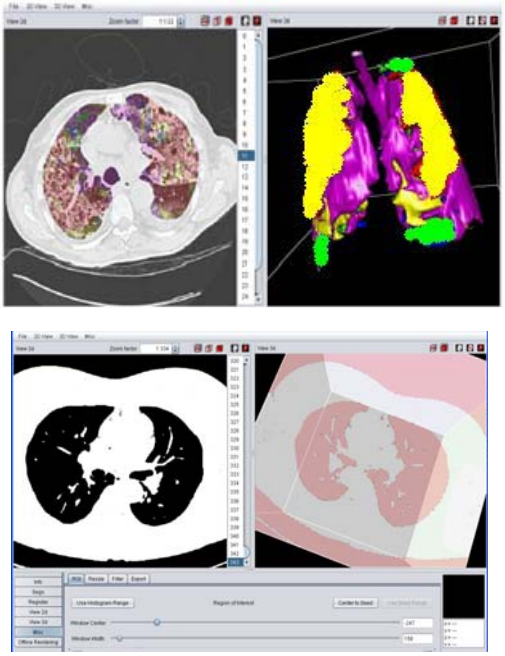
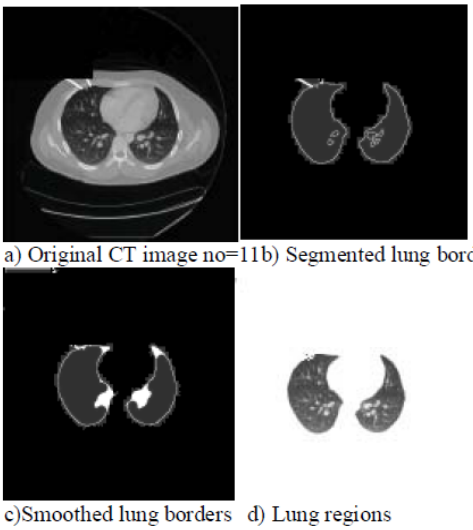


Figure 8 (a), (b): Lung segmentation from chest CT scan using YaDiv

As YaDiv is still under testing phase, so needs more algorithms to be embedded to make it more users friendly and fast.

d. Segmentation of Lung using MATITK:

ITK is an open source medical imaging processing library written in C++. While MATLAB also has many medical imaging algorithms, it is nice to be able to make use of the algorithms available in ITK. Precisely for this purpose, MATITK is written, allowing users to access certain ITK algorithms in MATLAB. With the help of MATITK, biomedical image computing researchers familiar with MATLAB can harness the power of ITK algorithms while avoiding learning C++ and dealing with low-level programming issues. We have implemented the volumetric segmentation by using the MATLAB environment. Figure 9 (a)-(d) gives the result of watershed segmentation of slice no=10 using MATITK. Though MATITK can be used only on 3D images [18], but in Figure 9, only slice no 11 is shown with its segmentation. Using VIEW3D function we can see the segmentation of all the slices. It is evident through observation that the proposed system produces much smoother results than the schemes that have been used earlier. There is also no loss of lung nodules in this method. MATITK provides various other segmentation methods also like Levelset, Gardient Vector Flow and many more and all present comparable results. There is no mechanism available to see all the 3d slices in MATITK. View3d function has been used for the same as shown in Figure 10.



a) Original CT image no=11b) Segmented lung borders

c)Smoothed lung borders d) Lung regions

Figure 9. Watershed Segmentation of lung CT using MATITK for slice=11.

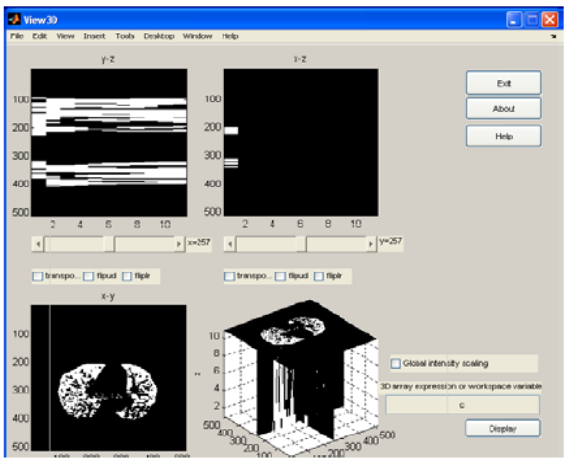


Figure 10. 3D view of MATITK results for ten slices.

B. *Automatic Segmentation of Lung CT*

For us, the goal is not to have a perfect segmentation but an algorithm that does not need manual intervention. It was not necessary to analyze all slices for 3D segmentation as our case database contains selected slices that represent certain pathology. On the other hand, there was no possibility to use information of connected slices to enhance the segmentation. The CT Images normally contains artifacts, noise which will not be suitable for further processing and hence it has to be pre-processed to reduce the noise. Following is the proposed technique which is applied to all the slices of chest CT to extract the lungs:

Step1: Image enhancement using filters to remove noise and thresholding.
Step2: Background removal.
Step3: Holes filling.

Step4: Get the original hounsefield values of lungs by bit Anding and multiplying with original image.

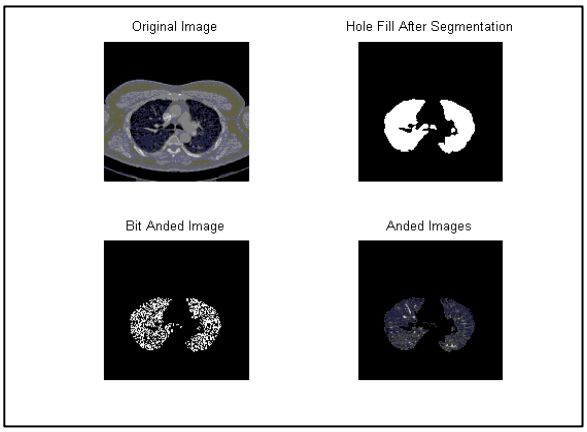


Figure 11. Lung segmentation from chest CT scan

Figure 11 shows the results of the proposed methodology. The results show that a seed point is required to start the segmentation and also the results vary depending upon the type of image in hand.

Figure 12 shows the pixel based segmentation on lung CT data. Various texture features have been extracted and compared with already saved features in the database for retrieval purposes.

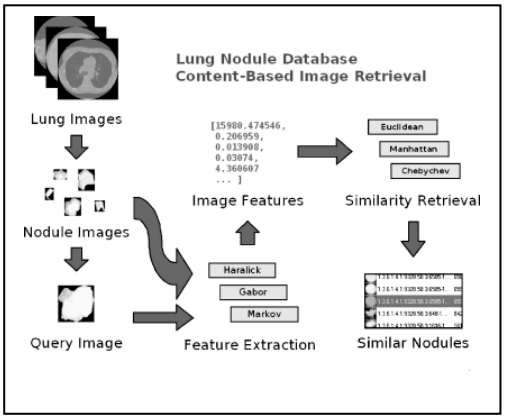


Figure 12. Automatic pixel based segmentation for lung nodule detection

V. Segmentation Methods For Thyroid Ultrasound

Various modality images can be used for detection of thyroid diseases [19]. However, the interpretation of US images, as performed by the experts, is still subjective. An image analysis scheme for computer aided detection of thyroid nodules would contribute to the objectification of the US interpretation and the reduction of the misdiagnosis rates. US imaging is currently the most popular diagnostic tool. It is inexpensive and easy to use. Before discussing the segmentation techniques for thyroid US images, it's necessary to understand its structure in details.

Structure of Thyroid Glands

The thyroid gland is a butterfly shaped organ and is composed of two cone-like lobes as shown in Figure 13. The organ is situated on the anterior side of the neck, lying against and around the larynx and trachea. It starts cranially at the oblique line on the thyroid cartilage (just below the laryngeal prominence or Adam's apple) and extends inferiorly to the fifth or sixth tracheal ring. It is difficult to demarcate the gland's upper and lower border with vertebral levels because it moves position in relation to these during swallowing.

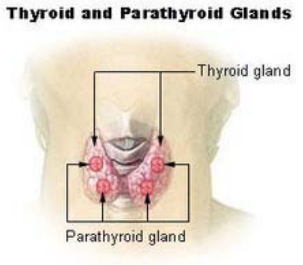


Figure 13. Structure of thyroid gland

Now, the segmentation of thyroid US images can be done in two ways.

1. Automatic Segmentation
2. Segmentation using Tools available

Both the methods are tried which will be briefed in the following sections and results were compared and analyzed.

A. Automatic Segmentation of thyroid US

There are various automatic segmentation techniques for thyroid US. Here we have mentioned simple techniques based on

1. Global thresholding
2. Local feature technique
3. Contrast enhancment

a . Global Thresholding Thyroid Region:

In a thyroid US image, the thyroid gland is always in the middle, below the bright part and above the dark part of the image. In [20], two reference values (R1 and R2) are defined to locate the probable thyroid region. R1 is the row index with the largest average intensity in the horizontal projection of the US image. R2 is the first row index with an average intensity of zero from the top to bottom in the horizontal projection of the US image. The probable thyroid region is located between the R1th row and the R2th row of the US thyroid image. An example of locating a probable thyroid region in an US thyroid image is shown in Figure 14.

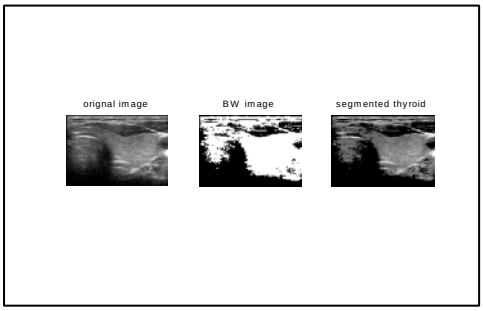
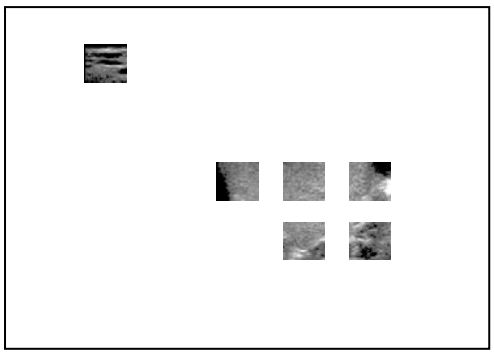


Figure 14. Probable Thyroid region



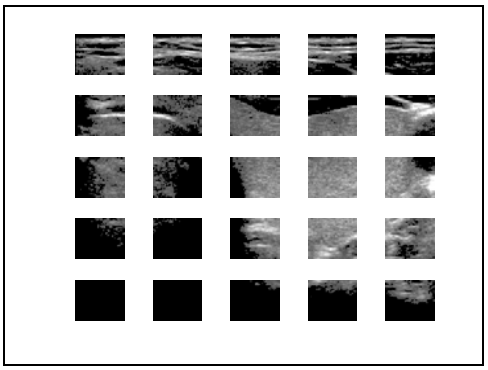


Figure 15(a): Image Divided into tiles
Figure 15(b): Tiles of ROI Extracted

b. Local feature technique:

[22]. Local features always give more accurate regions. Hence the image is divided into number of tiles and only the tiles of ROI are extracted from the image and rest of the tiles are rejected base on various statistical feature analysis as shown in Figure 15 (a)-(b). By increasing the number of tiles we can make the output very close to the actual thyroid shape but it becomes very much time consuming and it also increases the memory base for further analysis.

C. Thyroid US Segmentation using Automatic Tools

Ultra sound images of thyroid are not shape specific. The obvious reason is it is dependent on the angle by which image is taken. Different directions will produce different kind of images. Due to this segmentation becomes more critical. However texture of thyroid gland gives us very important information about nodules. Here two tools: Analyze 10.0 and Mazda are used for segmentation of thyroid US images as MATITK, YaDiv do not support 2D US images**.**

a. Segmentation of Thyroid US using Analyze:

Figure 16(a)-16(d) shows the output of using Analyze 10.0 on thyroid images.

b. Segmentation of US Thyroid using Mazda:

Figure 17 shows the result of segmentation using Mazda and also its features given by the tool. This tool helps in extracting the texture features which are very useful in finding the nature of nodules present in the thyroid glands.

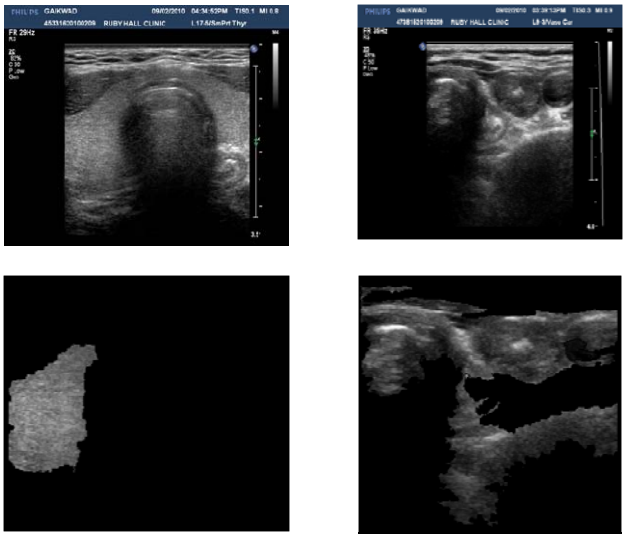


Figure 16 (a): US of normal Thyroid (b): US of abnormal Thyroid (c): Segmentation of normal Thyroid (d): Segmentation of abnormal Thyroid

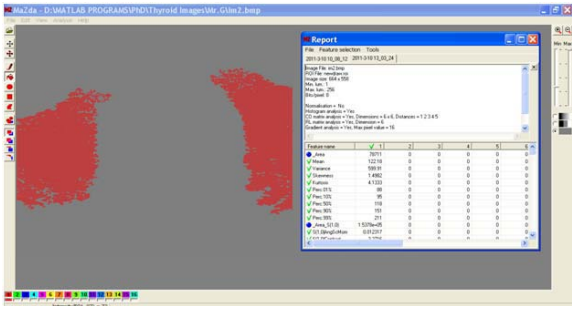


Figure 17: Segmented Thyroid and Features analysis

VI. Comparison Of All The Tools Used

The following results (Figure 18) shows the comparative analysis of all the results obtained either by automatic tools as well as by applying specific algorithm (automatic) segmentation on both lung CT as well as on thyroid US. From the results, it is concluded that automatic segmentation needs less manual intervention but needs more refinements like denoising and edge enhancements as well as 3D visualization problems whereas in automatic tools like Analyze , MazDa etc. no

such problems exist but a little manual interaction is required as well as the cost involved is more. These tools are actually are the research oriented tools but their commercial editions are also available and hence are used successfully in various hospitals like Mayo Clinic, Rochester, Minnesota.

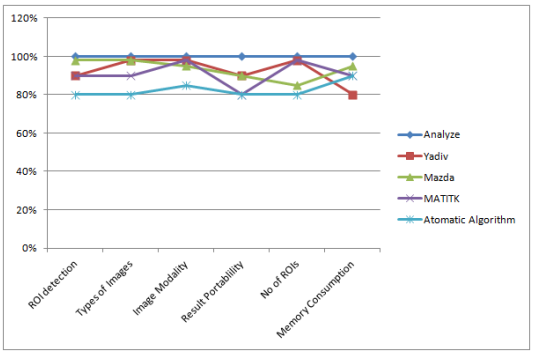


Figure 18: Comparative analysis of various tools

VII. Conclusion

Automatic tools actually make the segmentation task easier and flexible but these tools are commonly used in medical research and diagnosis. Moreover, the cost involved is more. Some of the tools like MATITK,YaDiv can be used only for 3D images like CT, MRI etc not for 2D like X-rays, US etc. Automatic segmentation method are little cumbersome and take more time and processing for segmentation. This paper basically provides a summary of existing automatic tools available to formulate the disease diagnosis part easier as well efficient. The future work is to develop new or improve the existing software tools to make the segmentation process easier and flexible for any modality.

References

[1] AM Hussain, G Packota, PW Major, C Flores-Mir, Role of different imaging modalities in assessment of temporomandibular joint erosions and osteophytes: a systematic review, Dentomaxillofacial Radiology (2008) 37, 63-71.

[2] Pradeep Singh; Sukhwinder Singh, Gurjinder Kaur (2008): A Study of Gaps in CBMIR using Different Methods and Prospective, Proceedings of world academy of science, engineering and technology, volume 36 , ISSN 2070-3740, pp. 492-496.

[3] Zhen Ma; João Manuel R. S. Tavares, R. M. Natal Jorge (2009): A review on the current segmentation algorithms for medical images, 1st International Conference on Imaging Theory and Applications (IMAGAPP), Lisboa, Portugal, INSTICC Press, pp. 135-140.

[4] S.S. Kumar, R.S. Moni, J. Rajeesh, "Automatic Segmentation of Liver and Tumor for CAD of Liver", Journal of advances in information technology, Academy Publisher, vol. 2, no. 1, February 2011, pp: 63-7

[5] Piotr M. Szczypiński, Michał Strzelecki, Andrzej Materka, Artur Klepaczko, " MaZda-A software package for image texture analysis", Elsevier, Computer Methods and Programs in Biomedicine, Volume 94 Issue 1, April, 2009, 66-76.

[6] K, K., & M, S. (2002). Patient-oriented Segmentation and Visualization of Medical Data. Computer, 214-219.

[7] [Online] http://www.analyzedirect.com/

[8] Vincent Chu and Ghassan Hamarneh, "MATLAB-ITK Interface for Medical Image Filtering, Segmentation and Registration", Medical Imaging 2006: Image Processing, Proc. of SPIE, Vol 6144, 61443T1-8.

[9] K. Awai and et al. Pulmonary nodules at chest CT: Effect of computer-aided diagnosis on radiologist's detection performance. Radiology, 230:347–352, 2004.

[10] K. G. Kim. Computer-aided diagnosis of localized ground-glass opacity in the lung at CT: Initial experience. Radiology, 237:657–661, 2005.

[11] P.Reeves, W. J. Kostis. Computer-aided diagnosis for lung cancer. Radiol. Clin. North Am., 38(3):497–509, 2000.

[12] Nisar Ahmed Memon, Anwar Majid Mirza, and S.A.M. Gilani, Segmentation of Lungs from CT Scan Images forEarly Diagnosis of Lung Cancer, World Academy of Science, Engineering and Technology 20 2006, pp: 113-118.

[13] Preeti Aggarwal, H.K. Sardana, Renu Vig, An Efficient Visualization and Segmentation of Lung CT scan Images for Early Diagnosis of Cancer, National Conference on Computational Instrumentation (NCCI-2010).

[14] Lew, John I; Rodgers, Steven E; Solorzano, Carmen C Developments in the use of ultrasound for thyroid cancer Current Opinion in Oncology: January 2010 - Volume 22 – Issue 1 - p 11-16.

[15] CE Engeler, JH Tashjian, SW Trenkner and JW Walsh, Ground-glass opacity of the lung parenchyma: a guide to analysis with high-resolution CT, American Journal of Roentgenology, Vol 160, 249-251.

[16] S. Tsantis, N. Dimitropoulos, D. Cavouras and G. Nikiforidis "A hybrid multi-scale model for thyroid nodule boundary detection on ultrasound images",Computer methods and Programs In Biomedicine, Volume 84, Issues 2-3, December 2006, Pages 86-98, Medical Image Segmentation Special Issue.

[17] K. G. Hollingsworth, D. J. Loma "Liver texture analysis: robustness of measurement in cirrhotic patients and healthy volunteers", Proc. Intl. Soc. Mag. Reson. Med. 13 (2005): 332.

[18] Holli et al., Texture analysis of MR images of patients with Mild Traumatic Brain Injury BMC Medical Imaging 2010, 10:8.

[19] Alexander S. Behnaz, James Snider, Chibuzor Eneh et.al, "Quantitative CT for Volumetric Analysis of Medical Images: Initial Results for Liver Tumors", Medical Imaging 2010, Proc. of SPIE Vol. 7623-76233U.

[20] Alison G Abraham, Donald D Duncan, Stephen J Gange, Sheila West "Computer-aided assessment of diagnostic images for epidemiological research" BMC Medical

Research Methodology 2009, 9:74doi:10.1186/1471-2288-9-74

[21] Chuan-Yu Chang, Yue-Fong Lei, Chin-Hsiao Tseng. Shyang-Rong Shih Thyroid Segmentation and Volume Estimation in Ultrasound Images, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 57, NO. 6, JUNE 2010.

[22] Neeraj Sharma, Lalit M Aggarwal, Automated medical image segmentation techniques, Journal of Medical Physics, Year: 2010, Volume: 35, Issue: 1, Page: 3-14.

Sonali Bhadoria, Assistant Professor, Electronics Department, MIT's Pune's MAE, Alandi, Pune, India. She has more than 10 years teaching experience. Presently she is pursuing her research in the area of cancer diagnosis using CBMIR techniques.
Email: sonali_nakade@yahoo.com.

Preeti Aggarwal, Assistant Professor in the Deptt. of Computer Sci. & Engg at Panjab University, Chandigarh, India. She has more than 11 years teaching and industrial experience. Presently she is pursuing her research in the area of cancer diagnosis using CBMIR techniques.
Email: pree_agg2002@yahoo.com.

Dr. C.G. Dethe, Principal, PIET,Nagpur. His field of specialization includes digital communication, Data network, Signal Processing. His papers are published in National and International Journals including IEEE proceedings. He is a fellow of IE and IETE and Life member of ISTE.
Email: cgdethe@yahoo.com

Dr. Renu Vig, Director, University Institute of Engineering and Technology, Panjab University, Chandigarh is an active member of IEEE and CSI. She has published more than 40 research papers. Her area of research is neural networks, fuzzy logic and signal processing.
Email: renuvig@hotmail.com.

Comparison Four Different Probability Sampling Methods based on Differential Evolution Algorithm

Lu Qingbo
College of Mechanical Electronic Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China
Email: jh0262@126.com

Zhang Xueliang and Wen Shuhua and Lan Guosheng
College of Mechanical Electronic Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China
Email: zhang_x_l@sina.com

***Abstract*—Differential Evolution (DE) is one kind of evolution algorithm, which based on difference of individuals. DE has exhibited good performance on optimization problem. The current studies almost are based on the simple random sampling method, and so this paper investigates other probability sampling methods, and proposed three novel differential evolution algorithms. The proposed algorithms are compared with the original differential evolution algorithm. The numerical results and Lorenz parameter estimation problem show that the new methods performed better than the original differential evolution algorithm.**

***Index Terms*—simple random sampling, stratified sampling, systematic sampling, cluster sampling, differential evolution, parameter estimation**

I. Introduction

Differential evolution (DE) is a stochastic, population-based optimization method[1, 2], which has been successfully to a wide range of problems as summarized in Price[3]. A number of variations of DE have been developed in the past decade to improve the performance. These researches can be divided two aspects, one is parameter investigation such as the mutation factor and the crossover probability [4-8], and the other is theoretical analyses [9-14]. Tvrdik[4] provided an experimental comparison of two different self-adaptive patterns and influence of exponential crossover. Das et al.[5] provided two new improved variants of DE, DE with random scale factor and DE with time varying scale factor. Liu and Lampinen[6] introduced a new version of the Differential Evolution algorithm with adaptive control parameters – the fuzzy adaptive differential evolution algorithm, which used fuzzy logic controllers to adapt the search parameters for the mutation operation and crossover operation. Teo[7] presented a first attempt at self-adapting the population size parameter in addition to self-adapting crossover and mutation rates. Lu et al.[8] proposed a modified differential evolution by randomly initializing and calculating the scale factor by chaos each generation and introducing a disaster factor into differential evolution algorithm. Omran[9] proposed barebones differential evolution algorithm, BBDE, which is a hybrid algorithm by capitalizing on the strengths of both the barebones PSO and self-adaptive DE strategies. Muelas et al.[10] combined the explorative/exploitative strength of the memetic algorithm and differential evolution algorithm and proposed a hybrid algorithm. Wang et al.[11] presented a novel Differential Evolution (DE) algorithm, called DE enhanced by neighborhood search (DENS), which differs from pervious works of utilizing the neighborhood search in DE, such as DE with neighborhood search (NSDE) and self-adaptive DE with neighborhood search (SaNSDE). Zhang and Sanderson [12] proposed an analytical method to study the evolutionary stochastic properties of the population in differential evolution (DE) for a spherical function and developed the properties of mutation and selection based on which a Gaussian approximate model of DE Zhang and Sanderson[13] proposed a new differential evolution algorithm, JADE, which proposed a new mutation strategy "DE/current-to-pbest" with the optional archive and updated control parameters in an adaptive manner. Zhang et al. [14] proposed a center differential evolution algorithm with adaptive crossover factor; the new algorithm incorporated the center point of the population into the DE algorithm.

Manuscript received August 23, 2011; revised December 5, 2011; accepted December 5, 2011.
Corresponding author, Zhang Xueliang.

doi:10.4304/jait.3.4.206-214

Simple random sampling method is used in all these researches. In this paper, we investigate the other three probability sampling methods and the three novel differential evolution algorithms are presented by applying these sampling methods into differential evolution algorithm. The proposed algorithms are compared with the original differential evolution algorithm. The numerical results and Lorenz parameter estimation problem show that the new methods performed better than the original differential evolution algorithm.

The reminder of the paper is organized as follows: four different probability sampling methods are summarized in Section 2. Section 3 summarized the differential evolution algorithm. The three novel different evolution algorithms are presented in Section 4. Section 5 presents the numerical results and discussions. Parameter estimation for the Lorenz system is investigated in Section 6. Finally, Section 7 concludes the paper.

II. FOUR DIFFERENT PROBABILITY SAMPLING METHODS

Probability sampling is a sampling technique wherein the samples are gathered in a process that gives all the individuals in the population equal chances of being selected. Probability sampling method has four different types, simple random sampling, stratified sampling, systematic sampling and cluster sampling.

A. Simple Random Sampling

Simple random sampling is the easiest form of probability sampling. This sampling method refers to a sampling method that has the following properties.

(a)The population consists of N individuals.

(b)The sample consists of n individuals.

(c)All possible samples of n individuals are equally likely to occur.

One of the best things about simple random sampling is the ease of assembling the sample. It is also considered as a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected.

B. Stratified Sampling

Stratified sampling is a probability sampling technique wherein the researcher divides the entire population into different subgroups (called strata), then randomly selects the final individuals proportionally from the different subgroups.

The strata do not overlap, as shown in figure 1, and they constitute the whole population so that each sampling unit belongs to exactly one stratum.

$$N = N_1 + N_2 + ... + N_H \quad (1)$$

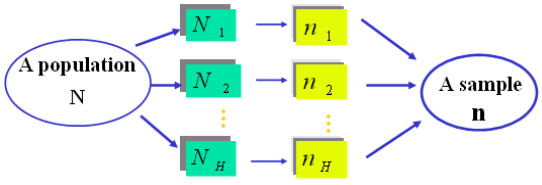


Figure 1. Stratification

If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. Stratified sampling consists of following steps:

(a)The entire population is divided into distinct subpopulations.

(b)Within each stratum, a separate sample is selected.

(c)Separate stratum means (or other statistics) are computed and then properly weighted to form a combined estimate for the entire population.

(d)The variances are computed separately within each stratum and then properly weighted and added into a combined estimate for the population.

C. Systematic Sampling

Systematic sampling is a random sampling technique which is frequently chosen by researchers for its simplicity and its periodic quality. In systematic sampling, the researcher first randomly picks the first item or subject from the population. Then, the researcher will select each n'th subject from the list.

For example, the researcher has a population total of 100 individuals and need 12 subjects. He first picks his starting number 5.Then the researcher picks his interval, 8. The members of his sample will be individuals 5, 13, 21, 29, 37, 45, 53, 61, 69, 77, 85, 97.

D. Cluster Sampling

In cluster sampling, instead of selecting all the subjects from the entire population right off, the researcher takes several steps in gathering his sample population.

First, the researcher selects groups or clusters, and then from each cluster, the researcher selects the individual subjects by either simple random or systematic random sampling. The researcher can even opt to include the entire cluster and not just a subset from it.

For example, a researcher wants to survey academic performance of high school students in China.

(a)He can divide the entire population (population of China) into different clusters (cities).

(b)Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

(c)Then, from the selected clusters (randomly selected cities) the researcher can either include all the high school students as subjects or he can select a number of subjects from each cluster through simple or systematic random sampling.

E. Comparison of the Four Probability Sampling Methods

Stratified sampling offers several advantages over simple random sampling. A stratified sample can provide greater precision than a simple random sample of the same size. Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money. We can ensure that we obtain sufficient sample points to support a separate analysis of any subgroup. The main disadvantage of a stratified sample is that it may require more administrative effort than a simple random sample. The main disadvantage of cluster sampling is that cluster sampling generally provides less

precision than either simple random sampling or stratified sampling at the sample size. When the increased sample size is sufficient to offset the loss in precision, cluster sampling may be the best choice. Systematic sampling is to be applied only if the given population is logically homogeneous, because systematic sample units are uniformly distributed over the population.

III. Differential Evolution Algorithm

Differential evolution (DE) is an evolutionary algorithm proposed by Storn and Price. The basic DE algorithm is described in detail below with reference to the four key operators: initialization, mutation, crossover and selection.

Initialization: Before the population can be initialized, both upper and lower bounds for each parameter must be specified. Once initialization bounds have been specified, a random number generator assigns each parameter of every vector a value from within the prescribed range. For example, the initial value (g=0) of the j^{th} parameter of the i^{th} vector is

$$x_{i,j}(0) = \text{rand}(0,1)(b_{U,j} - b_{L,j}) + b_{L,j} \quad (2)$$

The random number generator, rand (0, 1), returns a uniformly distributed random number from within the range [0,1].

Mutation: Once initialized, DE mutates and recombines the population to produce a population of NP trial vectors. For each parent, $x_i(t)$, of generation t, a trail vector, $u_i(t)$, is created by mutating a target vector. The target vector, $x_{r3}(t)$, is randomly selected, with $i \neq r_3$. Then, two individuals $x_{r1}(t)$, and are randomly selected with $i \neq r_2 \neq r_1$, and the difference vector, $x_{r1}(t) - x_{r2}(t)$, is calculated. The trail vector is then calculated as

$$u_i(t) = x_{r3}(t) + F\left(x_{r1}(t) - x_{r2}(t)\right) \quad (3)$$

where the last term represents the mutation step size. In the above, F is a scale factor used to control the amplification of the differential variation. Note that $F \in (0,2)$.

Crossover: DE follows a discrete recombination approach where elements from the vector, $x_i(t)$, are combined with elements from the trail vector, $u_i(t)$, to produce the offspring, $v_i(t)$. Using the binomial crossover,

$$v_{i,j}(t) = \begin{cases} u_{i,j}(t) & if\ rand(0,1) < Cr\ or\ j = r \\ x_{i,j}(t) & otherwise \end{cases} \quad (4)$$

where $j = 1,2,...,D$ refers to a specific dimension. r= rand(0, D). In the above, Cr is the probability of reproduction (with $Cr \in [0,1]$).

Selection: DE evolution implements a very simple selection procedure. The generated offspring, $v_i(t)$, replaces the parent, $x_i(t)$, only if the fitness of the offspring is better than that of the parent.

Storn and Price also proposed ten different strategies for DE based on the individual being perturbed, the number of individuals used in the mutation process and the type of crossover used. The strategy described above is known as DE/rand/1, meaning that the target vector is randomly selected, and only one difference vector is used. This strategy is considered to be the most widely used and it is the one used in this paper. Other main strategies include DE/best/1, DE/best/2, and DE/rand-to-best/1. The notation, DE/*x*/*y*, is used where *x* represents the individual being perturbed and *y* is the number of difference vectors used to perturb *x*.

IV. Three Novel Differential Evolution Algorithm

In original differential evolution algorithm, the method used to generate trail vector and target vector is simple random sampling method. The mainly different between the three novel differential evolution and DE is that the method used in the three novel differential evolution algorithms to generate trail vector and target vector are not randomly selected, but stratified sampling method, systematic sampling method and cluster sampling method.

A.Stratified Sampling Differential Evolution Algorithm

The stratified sampling differential evolution algorithm (SSDE) uses the stratified sampling method to generate the trail vector and target vector. In order to use stratified random sampling method, a quick sorting method is employed to sort the population by fitness. NP represents the population size. The pseudo code of the stratified sampling process is listed in Algorithm 1.

Algorithm 1 The pseudo code of the stratified sampling process

```
getIndex (int* r1,int* r2,int* r3,int i){
  Quick sort population by fitness;
  //select the layer index
  int k1,k2,k3=rand(1,3) and k1≠k2≠k3;
  if(k1==1 and k2==2 and k3==3)
    *r1=rand(0,[NP/3]),* r2=rand([NP/3], [2NP/3]),
    *r3=rand([NP/3],NP]) and *r1 or *r2 or *r3≠i;
  if(k1==1 and k2==3 and k3==2)
    *r1=rand(0,[NP/3]), *r2=rand([NP/3],NP]) ,
    * r3=rand([NP/3], [2NP/3]) and *r1 or *r2 or *r3≠i;
  if(k1==2 and k2==1 and k3==3)
    * r1=rand([NP/3], [2NP/3]),*r2=rand(0,[NP/3]),
    *r3=rand([NP/3],NP]) and *r1 or *r2 or *r3≠i;
  if(k1==2 and k2==3 and k3==1)
    * r1=rand([NP/3], [2NP/3]), *r2=rand([NP/3],NP]) ,
    *r3=rand(0,[NP/3]) and *r1 or *r2 or *r3≠i;
  if(k1==3 and k2==1 and k3==2)
    *r1=rand([NP/3],NP]), *r2=rand(0,[NP/3]),
    * r3=rand([NP/3], [2NP/3]) and *r1 or *r2 or *r3≠i;
```

```
1        if(k1==3 and k2==2 and k3==1)
3
1      *r1=rand([NP/3],NP]) , * r2=rand([NP/3], [2NP/3]),
4      *r3=rand(0,[NP/3]) and *r1 or *r2 or *r3≠i;
1      }
5
```

B.Systematic Sampling Differential Evolution Algorithm

The systematic sampling differential evolution algorithm (SYSDE) uses the systematic sampling method to generate the trail vector and target vector. The pseudo code of the systematic sampling process is listed in Algorithm 2.

Algorithm 2 The pseudo code of the systematic sampling process

```
   getIndex (int* r1,int* r2,int* r3,int i){

0        *r1=rand(0,[(NP-NP%3)/3]) and *r1≠i;
1
0        *r2=*r1+(NP-NP%3)/3;
2
0        *r3=*r2+(NP-NP%3)/3;
3
0     }
4
```

C.Cluster Sampling Differential Evolution Algorithm

The cluster sampling differential evolution algorithm (CDE) uses the cluster sampling method to generate the trail vector and target vector. The pseudo code of the cluster sampling process is listed in Algorithm 3.

Algorithm 3 The pseudo code of the cluster sampling process

```
   getIndex (int* r1,int* r2,int* r3,int i, int NC ){

0     //randomly select the subgroup index.
1     t=rand(0,NC);
0     //calculate the number of individual in each subgroup.
2     count=(NP-NP%NC)/NC;
0     //calculate the start number of the selected subgroup.
3     startNum=(NP-NP%N)/N*temp;
0      endNum=startNum+count;
4
0      *r1=rand(startNum,endNum), and *r1≠i;
5
0     *r2=rand(startNum,endNum), and *r2≠*r1≠i;
6
0     *r3=rand(startNum,endNum), and *r3≠*r1≠*r2≠i;
7
0     }
8
```

V. EXPERIMENTAL RESULTS

This section compares the performance of the SSDE algorithm, SYSDE algorithm and CDE algorithm with the original differential evolution algorithm (SDE).

The following functions have been used to compare the performance of SSDE, SYSDE and CDE with SDE. These benchmark functions provide a balance of unimodal and multimodal functions.

A .Sphere function, defined as

$$f(\boldsymbol{x}) = \sum_{i=1}^{D} x_i^2$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-100,100]$.

B. Schwefel' s problem 2.22, defined as

$$f(\boldsymbol{x}) = \sum_{i=1}^{D} |x_i| + \prod_{i=1}^{D} |x_i|$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-10,10]$.

C. Step function, defined as

$$f(\boldsymbol{x}) = \sum_{j=1}^{D} (\lfloor x_i + 0.5 \rfloor)^2$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-100,100]$.

D. Rosenbrock function, defined as

$$f(\boldsymbol{x}) = \sum_{j=1}^{D-1} \left[100\left(x_j^2 - x_{j+1}\right)^2 + \left(x_j - 1\right)^2 \right]$$

where $\boldsymbol{x}^* = (1,1,...,1), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-30,30]$.

E. Rotated hyper-ellipsoid function, defined as

$$f(\boldsymbol{x}) = \sum_{i=1}^{D} \left(\sum_{j=1}^{i} x_j \right)^2$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-100,100]$.

F. Generalized Swefel's problem 2.26, defined as

$$f(\boldsymbol{x}) = \sum_{j=1}^{D} -x_i \sin\left(\sqrt{|x_i|}\right)$$

where $\boldsymbol{x}^* = (420.9678, 420.9678, ..., 420.9678)$

$f(\boldsymbol{x}^*) = -12569.5$ for $x_i \in [-500,500]$.

G. Rastrigin function, defined as

$$f(\boldsymbol{x}) = \sum_{j=1}^{D} \left[x_i^2 - 10\cos(2\pi x_i) + 10 \right]$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-5.12,5.12]$.

H. Ackley's function, defined as

$$f(\boldsymbol{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2} \right) - \exp\left(\frac{1}{D}\sum_{i=1}^{D} \cos(2\pi x_i) \right) + 20 + e$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-32,32]$.

I. Griewank function, defined as

$$f(\boldsymbol{x}) = \frac{1}{4000}\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$$

where $\boldsymbol{x}^* = (0,0,...,0), f(\boldsymbol{x}^*) = 0$

for $x_i \in [-600, 600]$.

J. Six-hump Camel-back function, defined as

$$f(\boldsymbol{x}) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$$

where $\boldsymbol{x}^* = (-0.08983, 0.7126)$

$f(\boldsymbol{x}^*) = -1.0316285$ for $x_i \in [-5, 5]$.

Sphere, Schwefel's problem 2.22, Rosenbrock and Rotated hyper-ellipsoid are unimodal, while the Step function is a discontinuous unimodal function, Schwefel's problem 2.26, Rastrigin, Ackley and Griewank are difficult multimodal functions where the number of local optima increase exponentially with the problem dimension. The Camel-back function is a low-dimensional function with only a few local optima.

For all the algorithms used in this section, the population size NP set 100. All functions were implemented in 30 dimensions except for the two-dimensional Camel-back function. The results reported in this section are averages and standard deviations over 50 simulations. Each simulation was allowed to run for 50,000 evaluations of the objective function. *F*=0.5 and *Cr*=0.1, used DE/rand/1/bin strategy.

Table 1 summarizes the results obtained by applying the different approaches to all benchmark functions. The results show that the CDE, SSDE and SYSDE performed better than DE. The SSDE performed best than other three approaches. Fig.2 illustrates results for the selected benchmark functions. For the Sphere function, Fig.2a shows that SSDE achieved a faster reduction in fitness than SYSDE, CDE and SDE. For the Rosenbrock function, Fig.2b shows that SSDE achieved a faster reduction in fitness than SYSDE, CDE and SDE, and reached a good solution faster than the other approaches. For the Rotated hyper-ellipsoid function, Fig.2c shows that SSDE reached a good solution than other algorithms. For the Schwefel problem 2.26 function, Fig.2d shows that SSDE reached a good solution faster than the other approaches. For the Rastrigin function, Fig.2e shows that SDE achieved a faster reduction in fitness than the other approaches, but SSDE and CDE obtained a good solution than the other approaches. For the Ackley function, Fig.2f shows that SSDE achieved a faster reduction in fitness than the other approaches.

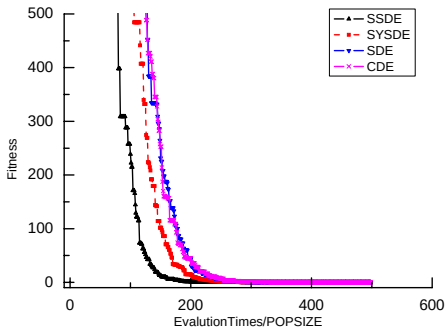
TABLE I.
MEAN AND STANDARD DEVIATION OF THE BEST-OF-RUN SOLUTION FOR 50 RUNS

Function	SDE	SYSDE	CDE	SSDE
Sphere	0.001157 (0.000253)	0.000072 (0.000018)	0.000725 (0.000324)	**0.00000 (0.00000)**
Schwefel Problem2.22	0.00491 (0.00055)	0.00129 (0.00017)	0.00367 (0.00074)	**0.00004 (0.00001)**
Step	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.00000 (0.00000)
Rosenbrock	151.788 (23.856)	73.267 (19.6488)	144.566 (36.9609)	**40.644227**
Rotated hyper-ellipsoid	19786.739 (2888.44)	18660.5212 (2485.94)	19789.3976 (3230.92)	**15844.9244 (2219.54)**
Schwefel problem 2.26	11509.472 (312.0277)	-12315.330 (206.56534)	-12168.749 (256.6817)	**-12569.483 (0.00364)**
Rastrign	40.760631 (4.099100)	37.562365 (3.180995)	36.836295 (4.488318)	**5.76638 (3.06823)**
Ackley	0.009931 (0.001264)	0.002387 (0.000298)	0.007937 (0.001949)	**0.00011 (0.00002)**
Griewank	0.009259 (0.00274)	0.001280 (0.00097)	0.006949 (0.0048)	**0.00001 (0.000001)**
SixJump	-1.031628 (0.000000)	-1.031628 (0.000000)	-1.031628 (0.000000)	-1.031628 (0.000000)

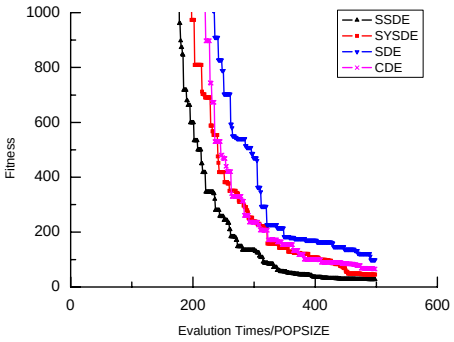
Figure.3 illustrates diversity for selected benchmark functions. Diversity has been calculated using

$$diversity = \frac{1}{POPSIZE}\sum_{i=1}^{POPSIZE}\sqrt{\sum_{j=1}^{D}\left(x_{ij}(t) - \bar{x}_j(t)\right)^2}$$

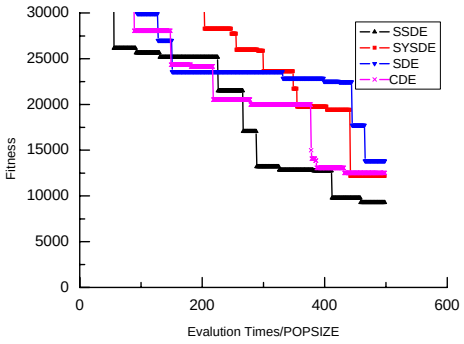
where $\bar{x}_j$ is the average of the *j*th dimension over all individuals, i.e. $\bar{x}_j(t) = \frac{1}{POPSIZE}\sum_{i=1}^{t} x_{ij}(t)$.



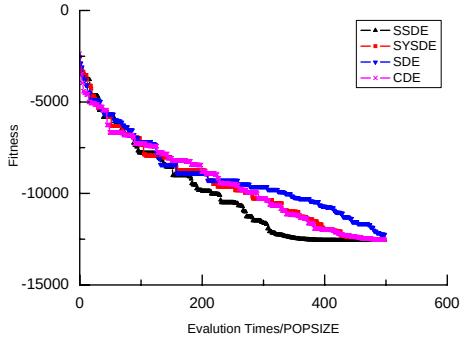
(a)Sphere(zoomed)



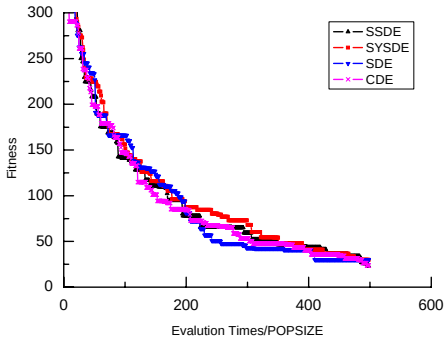
(b)Rosenbrock(zoomed)



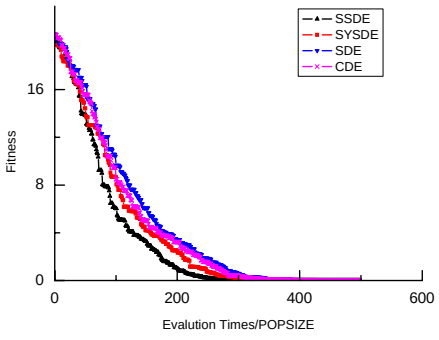
(c)Rotated hyper-ellipsoid(zoomed)



(d)Schwefel problem 2.26



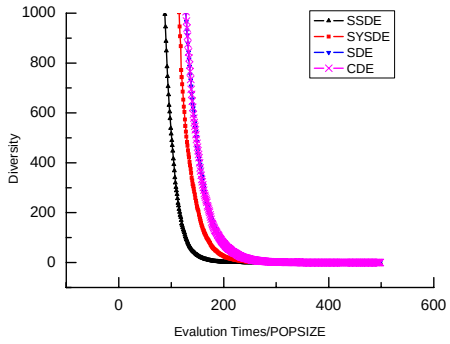
(e) Rastrigin(zoomed)



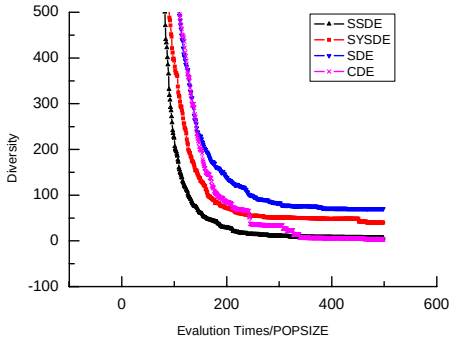
(f) Ackley

Figure 2. Performance comparison of the different methods for selected benchmark functions

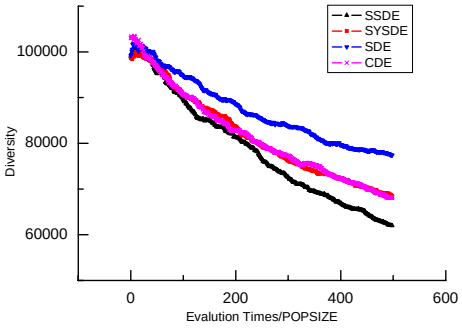
For the Sphere function, Fig.3a shows that the SSDE exhibited the fasted reduction in diversity enabling it to converge faster the other approaches. For the Rosenbrock function, Fig.3b shows that the SSDE exhibited the fasted reduction in diversity enabling to converge faster than the other approaches. The SDE was the slowest reduction in diversity which might cause its slow convergence. For the Rotated hyper-ellipsoid function, Fig.3c shows that SSDE achieved a faster reduction in diversity than other algorithms. For the Schwefel problem 2.26 function, Fig.3d shows that diversity increased firstly and then decreased, and SSDE exhibited the fasted reduction in diversity enabling to converge faster than the other approaches. For the Rastrigin function and Ackley, Fig.3e and Fig.3f show that SSDE exhibited the fasted reduction in diversity enabling to converge faster than the other approaches.



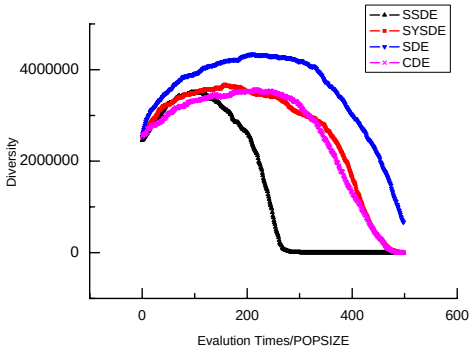
(a)Sphere(zoomed)



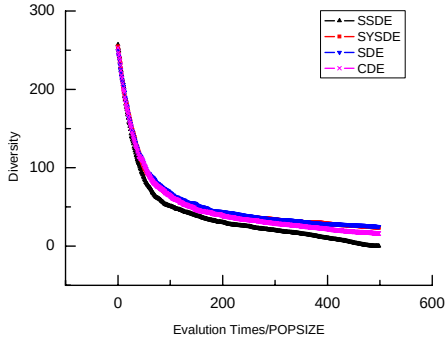
(b)Rosenbrock(zoomed)



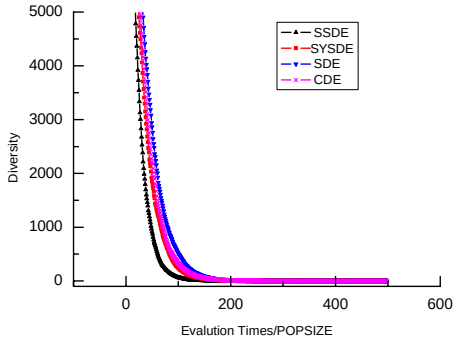
(c)Rotated hyper-ellipsoid



(d)Schwefel problem 2.26



(e) Rastrigin



(f) Ackley(zoomed)

Figure 3. Diversity comparison of the different methods for selected benchmark functions

VI. PARAMETER ESTIMATION IN THE LORENZ MODEL

The Lorenz model has been widely used for studies involving prediction and data assimilation in chaotic systems. The model consists of three variables x, y and z, which evolve according to the equations

$$x' = a(y - x) \tag{5}$$

$$y' = bx - y - xz \tag{6}$$

$$z' = xy - cz \tag{7}$$

where a,b,c are three constant parameters that are generally given the values 10, 28 and 8/3, respectively. For these values, the model variables follow a highly chaotic orbit. All calculations in this paper were performed using a fourth-order Runge-Kutta method with a time-step of 0.01.

The optimization model is established as follows:

$$\min\varepsilon = \frac{1}{M}\sum_{k=1}^{M}\left\|X_k - Y_k\right\|^2 \tag{8}$$

where M is the sequence length of state variable, X_k(k=1,2,…,M) is the kth state variables sequence at the true value of parameters of chaotic system, and Y_k (k=1,2,…,M) is the kth state variables sequence at the estimated value of parameters of chaotic system. For the Lorenz chaotic system, a, b, c are the decision variable.

For all the algorithms used in this section, the population size NP set 60. The results reported in this section are averages and standard deviations over 20 simulations. Each simulation was allowed to run for 30,000 evaluations of the objective function. F=0.5 and Cr=0.1, used DE/rand/1/bin strategy.

Table 2 summarizes the best solution obtained by applying different approaches. The results show that the SSDE obtained the best one of the four solutions. Table 3 summarizes the statistical results obtained by applying different approaches. The results show that SSDE performed better than the other methods in all means and standard deviation of parameters and best fitness.

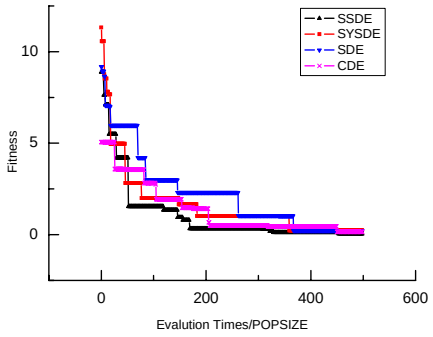
TABLE II.
COMPARISON OF THE BEST SOLUTION BY DIFFERENT METHODS

	SDE	SYSDE	CDE	SSDE
a	10.022700	9.926893	10.004860	10.002669
b	27.991236	28.062918	27.997970	27.993636
c	2.668656	2.667592	2.667006	2.665797
ε	0.219830	0.239681	0.149356	0.045487

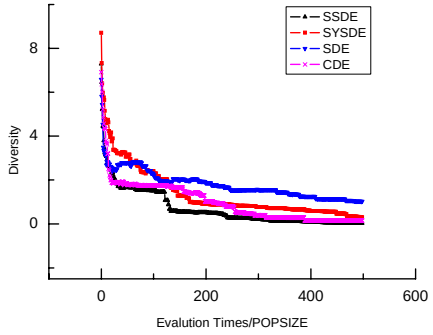
TABLE III.
STATISTICAL RESULTS OF DIFFERENT METHODS

	SDE	SYSDE	CDE	SSDE
a	10.06818 (0.23197)	10.03677 (0.17154)	10.0591 (0.21308)	10.01558 (0.03012)
b	27.93083 (0.19324)	27.95608 (0.14123)	27.94265 (0.20195)	27.98527 (0.03098)
c	2.66419 (0.00919)	2.66397 (0.00672)	2.66514 (0.01321)	2.66622 (0.00232)
ε	0.910464 (0.385202)	0.747560 (0.255342)	0.955321 (0.820303)	0.150204 (0.086611)

Figure 4 illustrates performance comparison of the different methods for Lorenz parameter estimation problem. Figure 4a shows that SSDE achieved a faster reduction in fitness and reached a good solution than the other approaches. Figure 4b shows that SSDE exhibited the fasted reduction in diversity enabling to converge faster than the other approaches.



(a)



(b)

Figure 4. Performance comparison of the different methods for Lorenz parameter estimation problem

VII. CONCLUSIONS

This paper presented three different models for differential evolution algorithm by investigating three probability sampling method. These approaches were compared with the simple random sampling method which used in the original differential evolution algorithm. The results show that these methods performed better than the original method in all selected benchmark functions. The results also show that the SSDE performed better than the SYSDE and CDE. This paper also investigated the parameter estimation problem, compared the results obtained by the proposed three algorithms and original differential evolution algorithm in the Lorenz parameter estimation problem. The results show that the SSDE performed best than the other strategies.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 50775153), Ph.D. Program Foundation of Ministry of Education of China (Grant No. 20091415110002).

REFERENCES

[1] Storn, R. and K. Price, "Differential Evolution-A simple and efficient adaptive scheme for global optimization over continuous spaces," *ICSI Technical Report TR-95-012,* March 1995.

[2] Storn, R. and K. Price, "Differential Evolution-a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 4, pp. 359-431, November 1997.

[3] Price, K. and R. Storn, "Differential Evolution: A Practical Approach to Global Optimization," *Springer*, 2005.

[4] Tvrdik, J., "Adaptive differential evolution and exponential crossover," *International Multiconference on Computer Science and Information Technology,* Wisia, 2008, pp.927-931.

[5] Das, S., A. Konar, and U.K. Chakraborty, "Two improved differential evolution schemes for faster global search," *Genetic and Evolutionary Computation Conference*, Washington, DC,2005, pp.991-998.

[6] Liu, J. and J. Lampinen, "A fuzzy adaptive differential evolution algorithm ," *Soft Computing - A Fusion of Foundations, Methodologies and Applications,* vol. 6, pp.448-462, September 2005.

[7] Teo, J., "Exploring dynamic self-adaptive populations in differential algorithm," *Soft Computing - A Fusion of Foundations, Methodologies and Applications,* vol. 8, pp.673-686, October 2006.

[8] Lu Qingbo, Zhang Xueliang, Wen Shuhua, et al., "Modified Differential Evolution and Its Application," *Transactions of the Chinese Society for Agricultural Machinery,*vol. 2, pp.193-197, February 2010.

[9] Omran, M.G.H., A.P. Engelbrecht, and A. Salman, "Bare Bones Differential Evolution," *European Journal of Operational Research,*vol.1, pp.128-139, July 2009.

[10] Muelas, S., A. La Torre, and J. Pea, "A Memetic Differential Evolution Algorithm for Continuous Optimization," *2009 Ninth International Conference on Intelligent Systems Design and Applications.* Pisa, 2009, pp. 1080 – 1084.

[11] Wang, H., Z. Wu, and S. Rahnamayan, "Differential Evolution enhanced by neighborhood search," *IEEE Congress on Evolutionary Computation.* Barcelona, 2010,pp. 1-8.

[12] Zhang, J. and A.C. Sanderson, "An approximate Gaussian model of differential evolution with spherical fitness functions," *IEEE Congress on Evolutionary Computation.* Singapore, 2007. pp. 2220-2228.

[13] Zhang, J. and A.C. Sanderson, "JADE: Adaptive Differential Evolution With Optional External Archive," *IEEE Congress on Evolutionary Computation.* NY,USA, 2009. pp. 945-958.

[14] Zhang Xueliang,Lu Qingbo, Wen Shuhua, et al., "Modified Differential Evolution for Constrained Optimization and Its Application" *Transactions of the Chinese Society for Agricultural Machinery,*vol. 8, pp.135-139, August 2008.

Lu Qingbo graduated from Northeastern University, China in 2000, and received the master of science degree from TaiYuan University of Science and Technology , China, in 2008. He is currently a Ph.D. candidate in TaiYuan University of Science and Technology ,China.

His research interests include mechanical designing and optimization.

Zhang Xueliang graduated from TaiYuan University of Technology, China in 1984, and received the Ph.D degree from Xi'an University of Technology, China, in 1998. He is currently a professor in TaiYuan University of Science and Technology ,China.

His research interests include mechanical designing and optimization and mechanical structural dynamics .

Wen Shuhua graduated from ShenYang University of Technology, China in 1987. She is currently a professor in TaiYuan University of Science and Technology ,China.

Her research interests include mechanical structural dynamics and optimization algorithm.

Lan Guosheng graduated from TaiYuan University of Technology, China in 2000, and received the master of science degree from TaiYuan University of Science and Technology , China, in 2005. He is currently a Ph.D. candidate in TaiYuan University of Science and Technology ,China.

His research interests include mechanical structural dynamics and optimization.

Nepenthes Honeypots Based Botnet Detection

Sanjeev Kumar, Rakesh Sehgal, Paramdeep Singh
Cyber Security Technology Division, C-DAC, Mohali, INDIA
ror.sanjeev@gmail.com, rks@cdacmohali.in, paramsbhatia@gmail.com

Ankit Chaudhary
Dept. of Computer Science, BITS Pilani, INDIA
ankitc.bitspilani@gmail.com

***Abstract*—the numbers of the botnet attacks are increasing day by day and the detection of botnet spreading in the network has become very challenging. Bots are having specific characteristics in comparison of normal malware as they are controlled by the remote master server and usually don't show their behavior like normal malware until they don't receive any command from their master server. Most of time bot malware are inactive, hence it is very difficult to detect. Further the detection or tracking of the network of theses bots requires an infrastructure that should be able to collect the data from a diverse range of data sources and correlate the data to bring the bigger picture in view.In this paper, we are sharing our experience of botnet detection in the private network as well as in public zone by deploying the nepenthes honeypots. The automated framework for malware collection using nepenthes and analysis using anti-virus scan are discussed. The experimental results of botnet detection by enabling nepenthes honeypots in network are shown. Also we saw that existing known bots in our network can be detected.**

***Index Terms*—Malware, Bots, Network Security, Nepenthes, Honeypots, Privacy**

I. INTRODUCTION

Botnet is becoming a major problem to internet as the size of cyber space is growing day by day. The applications running on the cyber space are becoming insecure and vulnerable to the attack generated by the black hat community. The motivation behind the attack can be to gain the access of the user's computer, to steal the information, to generate the DDoS attacks, to down the resources running in network. In last few years, the size of the network is increasing from low speed to gigabit network and the applications are also increasing day by day. In today's business oriented and heterogeneous network, security of the existing applications is extremely important. The flaws in security have become significant problems for private users, business, and even for government [1].

A botnet is a system that remotely controls malicious programs running on compromised hosts. Botnets are now a major source of network threats including DDoS, spam, identity theft, click frauds, etc. [17-19]. Botnets are still rapidly proliferating and communicating using a variety of protocols, such as IRC, HTTP, peer-to-peer, etc. The cumulative size of botnets is estimated in millions of hosts [16] [18-19]. Due to the huge number of botnets, and evolving botnet protocols, it appears difficult to block or remove (or both) all bots from the Internet. So, first start with minimum target to identify bots and their actions.

The attacks by Black hat community are daily increasing on the applications running in the network to steal useful information or to gain access of the client machine. Malwares are spreading into the cyber space and bot is one of those kinds of malware which has special kind of characteristics and remotely controlled by the botmaster. By detecting the some set of attacks used by the black hat community, we would be able to tighten the security of these kinds of high speed network as well as applications running in these networks.

As described in [16], bots are typically activated by bot commands through a communication and control channel (C&C) opened by attackers (i.e. botmasters) from remote sites. The bot commands issued may be run by a group of bots in the botnet simultaneously, as they have been programmed. The study of bot behavior in response to issued commands is important for the development of effective countermeasures, for tracing botnet growth, and for protecting the vulnerable infrastructures which are the target of bots. Also the identification of victims targeted by botnets may also be facilitated by a thorough analysis of bot commands [15]. More information about the bots and virus could be found out in [3-5].

Botnet detection and tracking is one of the very active research areas since last few years. Different solutions and techniques have been proposed for the same. Our approach was based on honeypot [7-8] for the malware collection and automated analysis of collected malware. A proactive system design has been discussed in [14]. The nepenthes low interaction honeypots was used instead of high interaction honeypots. There were two main reasons of using low interaction honeypots:-

- They can be easily configured and installed. Also they are low resource intensive.
- They are much faster than high interaction honeypots.

doi:10.4304/jait.3.4.215-221

So, with this consideration of honeypot technology, we could provide the attacker insight into our emulated network environment. Also we could monitor and log the interaction between the attackers and honeypots which could be further studied and analysed for botnet detection.

The system which was used to detect and analyse the attackers' action, was known as honeypot system. The honeypot has no intervention with the production traffic; therefore anything which comes on the honeypots is most likely the malicious intent. As compare to any other available security tools, honeypots are capable of logging much more information. They provide the vulnerable environment, to the attackers so that they can attack on the information system and can get access of the system.

Every activity of the attacker would be being monitored and logged and could be analysed. To tighten the security into the network, the nepenthes were deployed as a low interaction honeypots to study the known bot families into the network. Virtualization technology VirtualBox® [11] was used to reduce the involved hardware cost which is an open source virtualization product,which provides the flexible environment to set-up the network with single physical machine.

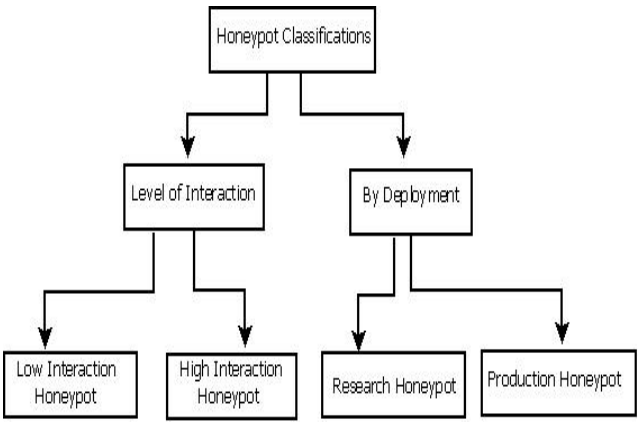


Figure1.Classification of honeypots

II. BACKGROUND ANDTECHNOLOGY USED

A brief introduction of different technologies is following which were used in the project.

A. Honeypot

A network security resource whose value lies in it, being scanned, attacked, compromised, controlled and misused by an attacker to attain his malicious goals. Lance Spitzner defines Honeypots as "A Honeypot is an information system resource whose value lies in unauthorized or illicit use of that resource" [1]. Honeypots can be classified into two main categories. Firstly, they can be based upon their level of interaction with an attacker. This can be further categorized as discussed in [9] [16]. Figure 1 depicts the classification of the honeypots according to their level of interaction and as per their deployment in network.

1) Low-interaction honeypots

Low interaction provides emulated network services to the attackers. Honeyd [6] and Nepenthes [13] are the examples of these kinds of low interaction honeypots. In contrast with low interaction honeypots, high interaction honeypots provides complete freedom to attackers to interact with real operating system and services and their all attempts are logged and accounted for.

2) Production Honeypots:

They are placed within an organization's production network for the purpose of detection. They extend the capabilities of intrusion detection systems. Such honeypots are developed and configured to integrate with the organization's infrastructure. They are usually implemented as low-interaction honeypots sitting within the server farm, but implementations may vary depending on requirements of the organization.

3) Research Honeypots:

These are deployed by network security researchers – the White hat hackers. They allow complete freedom for the attacker and learn their tactics in this process. Using research honeypots zero-day exploits, Worms, Trojans and viruses which are propagating in the network can be isolated and studied. Researchers can then document their findings and share them with system programmers, with network and system administrators, with various system and anti-virus vendors. They can provide the raw material for the rule engines of IDS, IPS and firewall systems.

B. Botnet

A 'Bot' has special characteristics as compare to the normal malwares. They are maintained and controlled by the remote servers known as botmasters. The collection of computers infected with such bot malwares are known as botnet. Therefore botnet is a network of zombies (infected computers) which are controlled by the botmaster. Normally bot malwares are inactive and get the command from the remote server (C&C). The commands are being executed by the bot client when it is given by the C&C servers. Bot masters control the botnet through a command and control mechanism. The formation of botnet is like C&C servers often communicate with other C&C servers to achieve the redundancy.

The topology of a botnet evolved over time from simple star to complex random combination of different topologies. Botnets are often classified according to the protocol through which it sends out commands to the zombie computers. A typical classification is as [2]:

- IRC Botnet: Bot masters acts as IRC servers and uses IRC channels to send commands to the botnet. All of the members of the botnet are connected to the channel. Commands are passed as a broadcast to the participants using the common IRC protocol.

- HTTP Botnet: Bot master acts as a web server and bots are connected to the web server. Commands are encapsulated in HTTP messages.
- P2P Botnet: Newer breed of botnet that uses existing P2P protocols to distribute commands. This kind of botnet is harder to detect compared to the other botnets.

The bots are connected to the botnet through a C&C channel as mentioned above. A C&C channel can operate on different network topologies and communication mechanisms. The most common protocol used for this is the IRC protocol. The main reasons why IRC is so popular are [12]:

- *Interactive*- the full two way communication between the server and the client is possible.
- *Easy to install-setting up private servers or using existing ones are easy.*
- *Easy to control-using credentials such as username, passwords and channels; all the needed functionalities already exist in the IRC protocol.*
- *Redundancy possibilities- by linking several servers together, one server can go down while the botnet is still functioning by connecting to other IRC servers.*

There also a botnet that uses the HTTP protocol for C&C. HTTP based C&C is still centralized, but the botmaster does not directly interact with the bots using chat like mechanisms. Instead, the bots periodically contact the C&C servers to obtain their commands. As its proven effectiveness and efficiencies, it is expected that centralized C&C (e.g. using IRC and HTTP) will still be widely used by botnets in near future.

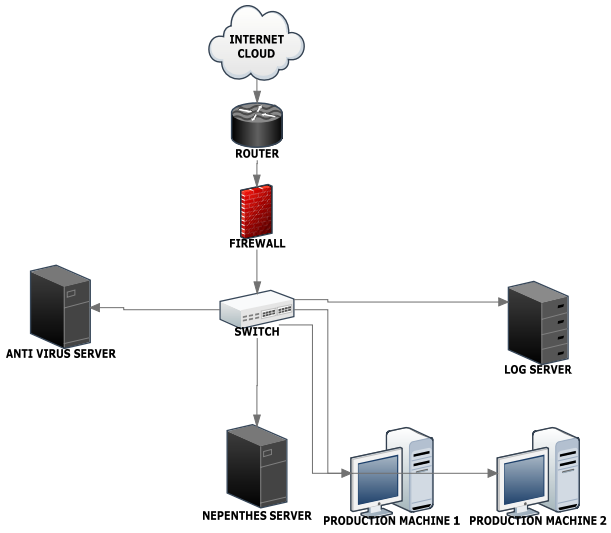


Figure 2.System Architecture

III. System Architecture

The nepenthes deployment architecture is discussed for malware collection and automated analysis using anti-virus scan. Nepenthes [13][15] are low interaction honeypots widely used for malware collection. Nepenthes as low interaction honeypot with default set of vulnerabilities which can be deployed in production network and can be useful to generate the alerts to system administrator who can take the necessary actions to tighten the security of the network. We enabled the nepenthes low interaction sensors in enterprise network and make them active all the time. We have deployed the nepenthes sensors in public network zone at various geographical locations as well as in private LAN to collect worms spreading in local private network [10].

Figure 2 show the system architecture including nepenthes and automated analysis of malwares using virus scan. As depicted in the Figure, there are nepenthes sensors used for malware collection. Basically there are three major components of the system: Malware collector, Virus scan server and Log server.

a) Malware Collector

As shown in the Figure 2, the low interaction honeypot (nepenthes) was used for malware collection. This is discussed how a particular low-interaction honeypot (nepenthes) [13] could be used to quickly alert an administrator about a network compromise. It captures malware and can assist in containing and removing the infection.

b) Nepenthes Sensors

For the implementation of nepenthes sensors VirtualBox® [11] was used. By using virtualization, it will reduce the hardware cost as compare to real physical system as well as will improve the deployment and maintenance.

Various Modules in Nepenthes

- *Vulnerability Modules* – emulates various services which look ripe for compromise to an attacker (lsass, dcom, veritas, dameware, etc)
- *Shell code Handlers and Emulators* – allows nepenthes to interact with the malware
- *Download Modules* – will download the binary (http, ftp, curl, etc)
- *Submission Modules* – will submit the binary for analysis (Norman, CWSandbox, postgres, etc)

c) Log server

Malware collected and all the data sets including network traces, pcap data were stored in log server for further analysis of the collected data. Log server is a central database server which keeps the metadata of the collected information. It keeps the following records:

- MD5 values of the malware samples
- Malware Binaries
- PCAP data & network traces
- Analysis results including antivirus labels etc
- IP address information
- Logs of the download, submitted binaries

Malware binaries stored in log servers were fetched and submitted to the analysis server where further analysis of the corresponding binary was done and result logs were putted into log server.

d) Anti-Virus Scan:

The malware binaries were fetched from log server and automatically submitted to anti-virus scan server which was doing the analysis of the binary based on predefined signatures. For this purpose three antivirus software were chosen from different companies MacAfee®, Symantec® and Microsoft®. Also the MD5 of the corresponding binary was submitted to the Virus Total [15] for scan with 42 anti-virus products. Virus Total is a free online service that enables Internet users to scan dubious files with 42 different antivirus (AV) tools.

The functionality of the system is as the following: the user sends a file to the system, via email or the web interface. He would get a report back when all AV tools will have finished examining the submitted file. That report includes the output of each engine, URLs with extra information about the potential threat if any. It gives information about file metadata size, various hashes of the file etc. It can also contain packet identification or the Portable Executable (PE) structure information of the malware. Virus Total with its 42 AV engines, offers a valuable service not only to the end users but also to the community of the AV vendors. Indeed, Virus Total can provide them with samples of malware that match certain criteria of interest to them. In the general case Virus Total sends a malware sample to AV vendor X then the following would be done:

- If at least one other AV engine has detected the sample as being malicious whereas the AV engine of X has not.
- If the AV engine from X has detected that sample as being malicious using a generic pattern or a heuristic.

Most AV vendors follow these two rules but some of them impose other criteria also. For instance, some have decided to get samples that are detected by at least N out of K AV engines and that their own has missed. Others do restrict even further the conditions by imposing that all engines from a well-defined subset of engines must have detected the sample and that their own has missed it [13]. Clearly the amount of samples to be sent to the AV vendors is a function of the filtering rules they have chosen. It is worth noting though that, in the general case, some vendors do get as many as 10000 samples per day [15].However this kind of malware collection mechanism may incur more cost and require complete collaboration with antivirus vendors in terms of services. This solution would be good for large organization, individual researchers, small organization, and private partners.

The complete process of malware collection and their analysis can be represented in following way:

1. If the nepenthes honeypots are deployed in public network zone, then there is central malware collection repository of the malware.
2. If the nepenthes honeypots are deployed in private network then there is local malware repository and analysis server

The system is working on 3 layer architecture. Layer 1 incorporates nepenthes honeypot sensors which captures the malware samples and sends data to the central server on a regular basis. Layer2 incorporates central server which performs activities like registering new nepenthes nodes, processing data sent by remote nodes by fusing the data with the configuration information of honeypots and converting the data in to a relational data base format. Layer 3 consist the database which acts as a data source for analysis engine. Figure 3 depicts the complete flow and deployment of the system.

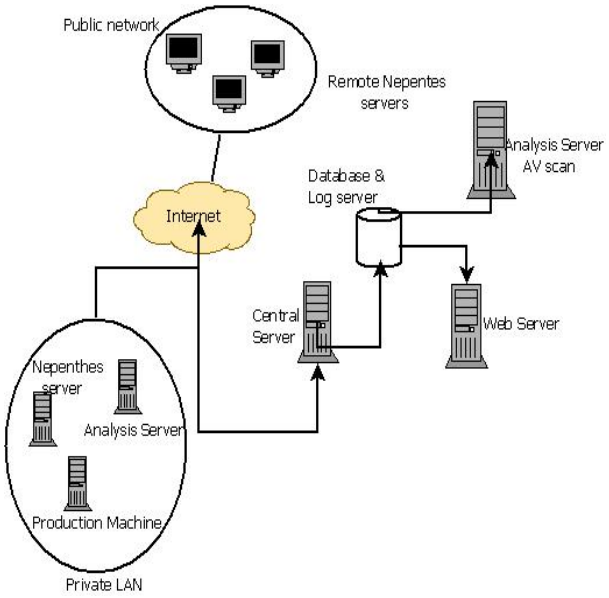
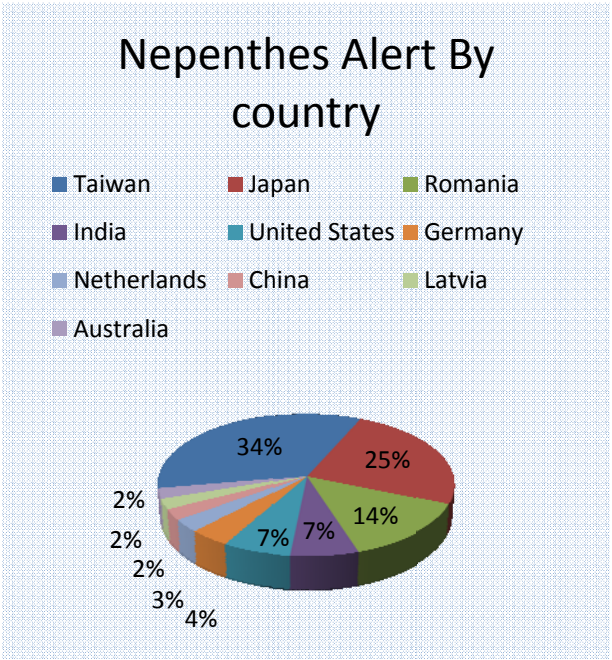


Figure 3.Complete Flow of the System

IV EXPERIMENTAL RESULTS

Here we are presenting some results of our nepenthes based system as malware collector and analysis mechanism. We have implemented our mentioned system on 1/1/2011 and collected very valuable information and malwares. We have collected real bots which were damaging the computers in network. During this period

of deployment total 732 numbers of samples were collected and we are containing large amount of PCAP data which is highly malicious in nature. Continuously we are submitting the data to our centre response team which are taking the remedial actions against the collected data sets and corresponding IP or attackers. Below Figure 4 illustrates the top 10 countries from where we have collected most our data sets.



Figures 4.Top 10 alerts from different countries

Below Table I show the top 10 most accessed URLs by deploying the nepenthes sensors geographically. Column 1 in Table represent the URLs accessed and column 2 represent the hit count of the corresponding URL. As shown in Table the hit count of the www.x.x.x.de/M.txt is 191 which are highest among others. For security concern, we have changed the name of the URLs.

TABLE I.
MOST ACCESSED URLS

URL	Count
http://www.x.x.x.de/M.txt	191
tftp://x.x.x.x/host.exe	97
http://y.net/fastenv	39
http://nmap.org/book/nse.html	24
http://y.proxyfire.net/fastenv	19
http://XX.63.156.12:8326/lsd	18
http://www.a.a.com	17
http://XXX.109.153.3/proxycheck.txt	15
http://XXX.109.153.5:11111	15
tftp://XXX.100.78.32/host.exe	14

TableII shows the top 10 MD5 values and their labelling corresponding to anti-virus scan. Column 1 represents the MD5 value of the collected malware binary, column 2 depicts hit count, and column 3, 4 and 5 depicts the antivirus labelling corresponding to MacAfee®, Microsoft®and Symantec® antivirus products. As we can see most of them are declared as IRC W32.IRC bot malwares. When we have submitted MD5 of that binary to Virus Total for scanning with 42 antivirus products, they were really the IRC bots and most of the antivirus products declaring them as IRC bots. So our deployed our automated system easily detected the bots spreading in the network and tighten the security against these bots.

TABLE II.
TOP 10 MD5 VALUES & THEIR LABELLING

MD5	Count	MacAfee	Microsoft	Symantec
865915650a85e7c27cdd11850a13f86e	51	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot
809fe9b32845edf5c09b871e0e68f227	63	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot
0da155b04f16dafafffbb1a485b3d0e1	27	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot
6e2fa9031a05b9649da062c550d14a3d	40	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot
f9dc3945bdd7406bd8db06a47963ec14	67	W32/Sdbot.worm.gen.bs	BackdoorWin32/Agent	
8a5ce07df6a5357dafa84f5317aaad35	75	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot
9019b23f2a5a51c33671739af2f30992	32	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot
15965bb88165d1eb06851d8f076130ba	31	W32/Sdbot.worm.gen.bs	BackdoorWin32/Rbot	W32.IRCBot

Also Figure 5 shows the daily collection graph by our system since its deploying date.

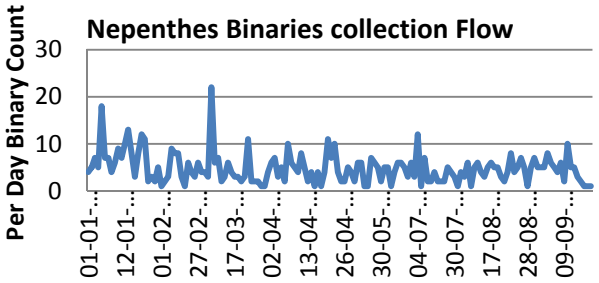


Figure5. Nepenthes Binaries Daily Collection

Some results of logs generated on the nepenthes honeypots are shown below when deployed with public IPs. Nepenthes honeypot IP address was 203.x.x.x and others were the foreign outside IP addresses. As we can see there is binary download ftp and tftp protocol. These results signify the interaction of the outside IP address with nepenthes as honeypots don't have any production values which clarify that they are malicious IP addresses. For security reason we have omitted the real IP address.

```
[2011-05-27T15:53:16]    66.x.x.x    ->
203.x.x.x creceive://66.x.x.x:9988/0
[2011-05-27T16:56:59]    59.x.x.x    ->
203.129.220.217
creceive://59.x.x.x:9988/0
[2011-05-28T04:57:15]    83.x.x.x    ->
203.x.x.x creceive://83.x.x.x:9988/0
[2011-05-28T12:03:57]    203.x.x.x   ->
203.x.x.x
tftp://203.111.222.65/host.exe
[2011-05-28T16:26:30]    203.x.x.x   ->
203.x.x.x tftp://203.x.x.x/host.exe
[2011-05-28T21:21:11]    60.x.x.x    ->
203.x.x.x creceive://60.x.x.x:9988/0
[2011-05-29T02:52:18]    203.x.x.x   ->
203.x.x.x
ftp://1:1@203.x.x.x:12405/host.exe
[2011-05-29T06:34:37]    203.x.x.x   ->
203.x.x.x tftp://203.x.x.x/host.exe
[2011-05-29T08:53:48]    209.x.x.x   ->
203.x.x.x creceive://209.x.x.x:9988/0
[2011-05-30T01:08:45]    174.x.x.x   ->
203.x.x.x creceive://174.x.x.x:9988/0
[2011-05-30T02:54:41]    125.x.x.x   ->
203.x.x.x http://www.baidu.com/
[2011-05-30T06:28:48]    203.x.x.x   ->
203.x.x.x tftp://203.x.x.x/host.exe
[2011-05-30T06:47:21]    203.x.x.x   ->
203.129.220.217
ftp://1:1@203.180.24.32:18807/host.exe
```

V. Conclusion And Future Work

In this paper we have presented an automated system based on nepenthes as malware collector and analysis of them using antivirus scan. This is one step of detecting the known bots in the network and in any organization we can detect the bot spreading in the network using this system. Further we have also shown the behavior based analysis of the collected malwares which are not detected by the antivirus products. The claim is that all the software (OS and associated tools) are Open Source. A low-interaction honeypot like nepenthes is easy to install and requires minimal maintenance. It may provide valuable information in the event of an infection within your organization. When used in conjunction with an Intrusion Detection System, valuable information about the behavior of the malware, packet captures and the malware binary itself may be obtained.

Acknowledgment

We would like to thank Cyber Security Technology team at C-DAC, Mohali to provide the infrastructure and recourses to collect the malwares and to available them for further analysis. We are also very thankful to Executive Director of CDAC, Mohali to provide us full support. This research was supported by DST, Ministry of Science & Technology, and Govt. of INDIA.

References

[1] "Security threat report: 2010", Sophos Group, 2010, DOI:http://www.sopos.com/security/topic/secuirtyreport-2010.html
[2] Microsoft Security BulletinMS03-026, "Buffer Overrun in RPC Interface Could Allow Code Execution".
[3] Description of the Blaster worm, DOI: www.symantec.com/security_response/writeup.jsp?docid=2003-081113-229-99.
[4] Description of the Mocbot/Wargbot worm, DOI: www.symantec.com/security_response/writeup.jsp?docid=2006-081312-3302-99.
[5] Spitzner, L. "Honeypots: Tracking Hackers", Addison Wesley, USA, 2002, pp. 1-430.
[6] Stoll, C., "The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage", Pocket Books, New York, 1990.
[7] The Honeynet Project,"*Know Your Enemy: Honeywall CDROM Roo",2005*Available: http://old.honeynet.org/papers/cdrom/Roo/index.html.
[8] Abbasi, F.H. and Harris,R.J., "Experiences with a Generation III virtual Honeynet", Telecommunication Network and Applications Conference (ATNAC), 2009.
[9] Wireshark,www.wireshark.org.
[10] Padmanabban,V. N.and Subramanian,L., "Determining the geographic location of Internet hosts", In SIGMETRICS/Performance, 2001, pp. 324– 325.
[11] VirtualBox. (2004). Sun VirtualBox® User Manual, Available: http://www.virtualbox.org/manual/UserManual.html Last accessed 20 July 2008.
[12] Stallings, W., "Cryptography and Network Security Principles and Practices", Third Edition, Prentice Hall, 2003.
[13] Edward, B. and Camilo, V.,"Towards aThird Generation Data Capture Architecture for Honeynets",Proceedings of the IEEE Workshop on Information Assurance and Security, USA, 2005, pp21-28.
[14] Chaudhary,A. and Raheja,J. L."A Formal Approach for Agent Based Large Concurrent Intelligent Systems",

International Journal of Advanced Engineering Technology, Vol. 1, June 2010, pp. 95-103.
[15] Barford, P. andYegneswaran, V., " An Inside Look at Botnets",Advances in Information Security, Springer, Vol. 27, 2007 pp. 171–191.
[16] Rajab, M. A., Zarfoss, J.,Monrose, F.and Terzis, A., "A MultifacetedApproach to Understanding the Botnet Phenomenon ", 6th ACM SIGCOMM conference on Internet measurement,ACM, 2006.
[17] Barford,P.and Yegneswaran,V., " An Inside Look at Botnets ", Special Workshop on Malware Detection, Advances inInformation Security, 2006.
[18] Rajab, M. A., Zarfoss, J., Monrose, F.andTerzis, A., "My Botnet is Bigger than Yours (Maybe, Better than yours): why size estimates remain challenging ", in First Workshop on Hot Topics in Understanding Botnets, 2007.
[19] Dagon, D., Zou, C. and Lee, W., " Modeling Botnet Propagation Using Time Zones",Network and Distributed SystemSecurity Symposium, The Internet Society, 2006.

Sanjeev Kumar received his B.Tech from Kurukshetra University, INDIA and pursuing M.Tech in CS from PTU INDIA. He is working as a staff scientist at CDAC, Mohali, INDIA. He is CCNA and CCNP certified and an active member of Indian National Grid known as GARUDA. His technical expertises are in networking, network security.

Rakesh Sehgal received his B.E in Electronics from Nagpur University, 1988 and M. Tech. in Computer Science from DAU, Indore. He is currently Principal Design Engineer & Head of Cyber Security Technology Division at CDAC- Mohali. He has vast research experience in Network Security, Honeynets and Honeypots.

Paramdeep Singh received his MCA from PTU Punjab INDIA. Currently he is working in Cyber Security Technologies Division at CDAC Mohali, INDIA. He has extensive work experience on Honeynets and Honeypots.

Ankit Chaudhary received his ME & PhD in CSE from Birla Institute of Technology & Science, Pilani, INDIA. His research interests are Machine Learning, Artificial Intelligence, Network Security and Mathematical Computations.

Web based Geo-Spatial and Village Level Information Extraction System using FOSS

Mr. Kodge B. G.
Department of Computer Science, S. V. College, Udgir, Dist. Latur (MH), India
kodgebg@hotmail.com

Dr. Hiremath P. S.
Department of Computer Science, Gulbarga University, Gulbarga (KA), India
hiremathps53@yahoo.co.in

***Abstract*— Increase in the number of commercial and open source tools for geospatial applications have resulted in confusing environment among students / researchers, institutions and consultants while selecting software tools. Also geospatial software applications require more time and money in design and development. To demonstrate the use of free open source software (FOSS) for geospatial data management, a system was designed for extracting geo-spatial information content from spatial database and practical issues during development are described in this paper.**

***Index Terms*— FOSS, GIS, Data extraction, Village information system.**

I. INTRODUCTION

In a developing country like India, the 73% of the population resides in rural area and cannot subscribe commercial software. Though number of tools are available for web based spatial information system, each one of them have some restrictions, such as cost, license, internet connectivity, availability of skilled person, etc. To overcome such problems, open source tools are better solutions. Open source geospatial tools provide facilities such as cost effectiveness and it can be customized according to user needs. Also, various spatial queries can be made for better decision making. They are helpful in monitoring various rural development schemes run by government [2].

The GIS provides a systematic spatial and attribute database, which is prerequisite for implementing development and research projects, for drawing development strategies that are sustainable, area specific and take into account the local needs; and to facilitate the process of decentralized (to smaller unit i.e. district or below) planning and for more timely response to promote effective administration, planning, decision making and development process. Study aims at understanding basics of open source software and further focusing on need and capabilities those software tools by taking case of web based information system.

II. OPEN SOURCE GEO-SPATIAL TOOLS

Geospatial Data: Almost all information to support rural development has a strong geographical context. Geospatial data include geographic coordinates (e.g., latitude and longitude) that identify a specific location on the Earth; and data that are linked to geographic locations or have a geospatial component (e.g., socio-economic data, land records, land surveys, homeland security information and environmental analyses).

FOSS are those software that have licenses that allow users to freely run the program for any purpose, modify the program as they want and also to freely distribute copies of either the original version or their own modified version.

A. Need for Open Source Geospatial Software Tools

The last 20 years have seen dramatic developments in GIS technology and geographical information science. High competition and growing user demand has resulted in a number of high-quality solutions, which are largely responsible for the vast increase in the GIS marketplace [1]. But the commercial software are not so much popular in small development projects because:

- Inadequate funds for uniform software setup.
- Poor support for commercial software packages.
- Diverging needs within one organization.
- There are a lot of restrictions for access to source code.
- Installation cost is too high, hence not economical for small project.
- User should have deep knowledge of the software.
- Inadequacy of standardized format for results (i.e., varying file and data formats).

B. Advantages of Open-source Geospatial Tools

- Access to source.
- Enables development of highly customized applications based on client's needs.
- Development priorities are driven by end-user needs.
- No licensing fees.

doi:10.4304/jait.3.4.222-227

- Resources are allocated for building the applications. No licensing multiple machines.
- Interoperability, adoption of open specifications.
- Developers listening to users directly.
- Issues can be resolved in-house.
- Affordable and high quality.
- Open source software has fewer defects, because if defects are present, they get repaired faster.
- Free as in freedom.

FOSS guarantees four fundamental freedoms, (i) To run the program for any purpose. (ii) To study how the program works, and adapt it to your needs. (iii) To redistribute copies. (iv) To improve the program and release your improvements to the public, so that the whole community benefits [3].

III. Desktop and web applications of open source geospatial tools

A web service can discover and invoke any service anywhere on the Web, independently of the language, location, machine, or other implementation details. The goal of Semantic Web Services is the use of richer, more declarative descriptions of the elements of dynamic distributed computation including services, processes, message-based conversations, transactions, etc. [21].

A. Natural Resource Data Base (NRDB)

NRDB Pro is a GIS tool for developing and distributing environmental databases. Its aim is to provide people in developing countries with a powerful yet simple tool to assist in managing their own resources.

The natural resources database software was originally developed for the Bohol Environment Management office, provincial government of Bohol, Philippines, by Richard D. Alexander, with the assistance of voluntary service overseas. This was supported through the skills for community-based resource utilization and management (SCRUM) program. SCRUM was a four-year project partly funded by the European Union (EU) and the British Embassy. The goal of the project was to promote effective resource management by communities so as to ensure food security and well-being, alleviate poverty and prevent further depletion of valuable natural resource in which their livelihoods depend [12].

B. Quantum GIS (QGIS)

Quantum GIS is a Geographic Information System that runs on Linux, Unix, Mac OS X, and Windows. The QGIS supports vector, raster, and database formats. It can access databases like PostGIS, in addition to the dozens of other vector and raster formats. It supports feature labeling and has a great user community. Extensibility is provided through a plugin environment [16].

C. *Integrated Land and Water Information System (ILWIS)*

The ILWIS is desktop GIS & Remote Sensing software, developed in the Netherlands by ITC up to its last release (version 3.3) in 2005. ILWIS software is available as open source software under the 52° North initiative (GPL license). Its powerful image processing functions make it a highly useful tool for natural resources management and for organizations that need to process orthophotos or satellite imagery for base mapping [9].

D. GeoServer

GeoServer is an open source software server written in Java that allows users to share and edit geospatial data. Designed for interoperability, it publishes data from any major spatial data source using open standards. Being a community-driven project, GeoServer is developed, tested, and supported by a diverse group of individuals and organizations from around the world. GeoServer is the reference implementation of the Open Geospatial Consortium (OGC), Web Feature Service (WFS) and Web Coverage Service (WCS) standards, as well as a high performance certified compliant Web Map Service (WMS). GeoServer forms a core component of the Geospatial Web [5].

E. POSTGRESQL

PostgreSQL is an object-relational database management system (ORDBMS) based on POSTGRES, Version 4.2, developed at the University of California at Berkeley Computer Science Department. POSTGRES pioneered many concepts that became available only in some commercial database systems much later.

PostgreSQL is an open-source descendant of this original Berkeley code. It supports a large part of the SQL standard and offers many modern features: complex queries, foreign keys, triggers, views, transactional integrity, multiversion concurrency control [15].

F. PostGIS

PostGIS adds support for geographic objects to the PostgreSQL object-relational database. In effect, PostGIS "spatially enables" the PostgreSQL server, allowing it to be used as a backend spatial database for geographic information systems (GIS), much like ESRI's SDE or Oracle's Spatial extension. PostGIS follows the OpenGIS "Simple Features Specification for SQL" and has been certified as compliant with the "Types and Functions" profile [14].

IV. How open source geospatial tools can be used for rural development?

Recent advances in the domain of spatial technology are making considerable impact in planning activities. This type of planning is more important in countries like India where rural population is more and is having variations in geographic patterns, cultural activities, etc. One of the strongest points in favor of open source geo spatial tools is that they are cost effective. By using open source geospatial tools, planning work can become simpler and cost effective. It is helpful in monitoring the rural development schemes and various spatial queries that can be run for analyzing the problem.

Various maps can be generated using open source geospatial tools, which provide an added dimension to data analysis by Geo-visualization.

- Administrative Map, Village Location Map, State Highways Map, Major District Roads.
- Map of Village Roads, Light Vehicle Roads (Matelled, Un-matelled), Roads Under construction, Prime Minister Roads Under Construction.
- Map showing Child and Maternity Welfare Centres, Ayurvedic, Allopathic, Homeopathic Hospitals, Community Health Centres, Public Health Centres, Base Hospital, District Hospital, Civil Hospital, Women Hospital etc.
- Perennial Water bodies Non Perennial Warter bodies of District, Forest Cover of District, Agriculture Land map, etc.
- Service area map, showing accessibility of service facility from villages or to villages.
- Map showing drought affected area, flood affected area can be prepared.

V. WEB BASED SPATIAL INFORMATION SYSTEM OF STUDY SITE

In order to demonstrate the use of open source tools for web based spatial information system was designed and tested on the stand alone platform and local area network. The state, district, taluka and village level data was obtained from Latur district spatial database. The NRDB (free open source software) is used to develop the spatial database.

A. Methodology

The tools selected were, NRDB for the geo-referencing and generating different outline layers from the maps. GeoServer for windows consists of set of Apache web server, etc. which were used for visualizing information over front end. POSTGIS is used to build spatial query module facilitating panning, zooming and selecting region to view the database in POSTGRESQL. Relational database management tool POSTGRESQL server was used to store data and PHP facilitated querying on database in the form of simple and complex non spatial queries. Open source content management system 'QGIS' was used to design the front end consisting of user account. The open source software used in this context are shown in Figure 1.

Figure 1. Process flow of spatial information system

B. Structure of the Proposed Information System

The internal logical structure of proposed information system is shown in Figure 2, which also shows interactivity with different parts of the system, and figure 3 shows the front end information systems. The complete documentation of the system is done in order to facilitate the next user in future to develop the better system than available.

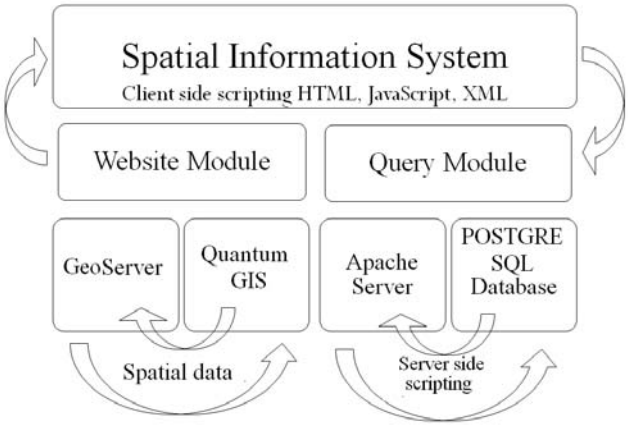


Figure 2. Logical Structure of Proposed Information System

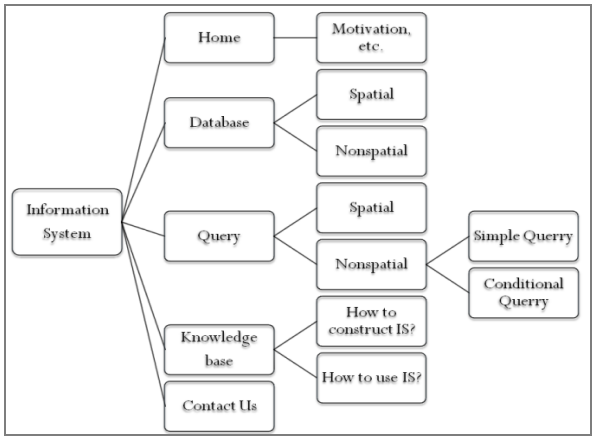


Figure 3. Front end of Information System

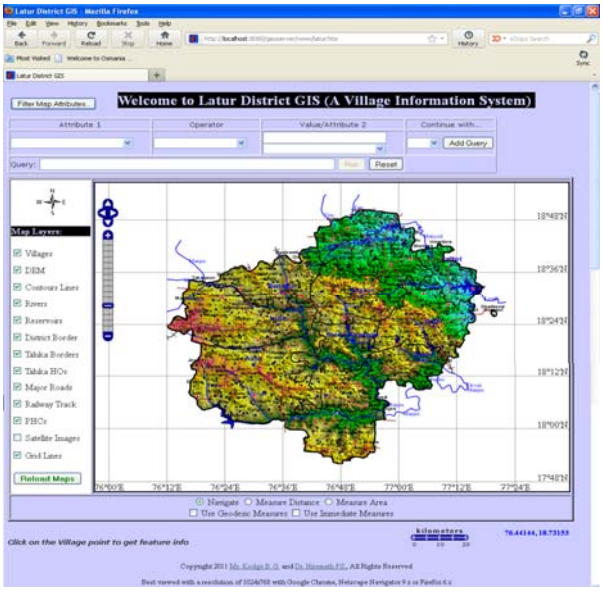


Figure 4. Web based Spatial Information System at http://localhost:8080

The proposed web based spatial information system is designed and implemented through Geoserver (web server) and tested at local network http://localhost:8080/ GeoServer port. This system provides map overlay

management through enabling or disabling of different loaded map layers like, village locations, digital elevation model (DEM), ASTER Satellite Imageries, Multicolored elevation contour line, etc. The layers are as shown in left panel (i.e Map Layers) of Figure 4. Apart from these, a user can also use some more features of this system i.e map navigation, zooming, scale manipulation, geolocation display, measuring distance, measuring area, etc.

C. Filter Map Attributes

Extraction of data with respect to user need is the key factor of this system. The above said system is designed for villages level information system, and allows users to extract any kind of village level information through filtering their attributes using common query language (CQL).

TABLE I.
LISTS OF ATTRIBUTES AND OPERATORS USED IN FILTERING MODULE.

Attribute1	Op	Attribute2	Logical Op.
Village Name House Holds Total Population Total Male-Population Total Female-Population Total Literates Total Male-Literates Total Female-Literates Total Percent-Literates Total Male-Percent Literates Total Female-Percent Literates Total Illiterates Total Male-Illiterates Total Female-Illiterates Primary Schools High Schools Colleges Primary Health-Centers Sub Centers Sex Ratio Village-Performance (%)	= = ! = < <= > >= LIKE	X-(User defined value) Total Male-Population Total Female-Population Total Literates Total Male-Literates Total Female-Literates Total Percent-Literates Total Male-Percent-Literates Total Female-Percent-Literates Total Illiterates Total Male-Illiterates Total Female-Illiterates Primary-Health-Centers Sub Centers	AND OR

The server side CQL are written to execute queries automatically through selecting the villages level attributes and operators which are provided in different list boxes in the filter attribute module (top left corner in Figure 4). The list of all attributes and operators of filter attribute module are shown in Table 1. By combining or selecting different attributes and their appropriate operators, user can extract the useful information from this system. Following are the few examples demonstrated to showcase the use of filter attribute module.

VI. RESULTS AND DISCUSSIONS

With respect to the above discussed concepts, the followings are few examples and results extracted from the proposed system.

Example 1:

Consider, a user would like to extract the information with multiple conditional expressions like, which villages are having more illiterates than literates and their sex ratio is below 1:1 (male to female).

The attribute filtering using logical AND, OR are used to execute multiple conditional expressions, and they can be processed using the fields provided in the filter module. This example will execute the query with following CQL statement.

New_Map = SELECT * FROM VILLAGES1 WHERE Totillit > TotLit AND SexRatio < 1 ;

The above query will extract the spatial locations of all villages which are satisfy the above query conditions, and they are shown in the Figure 5

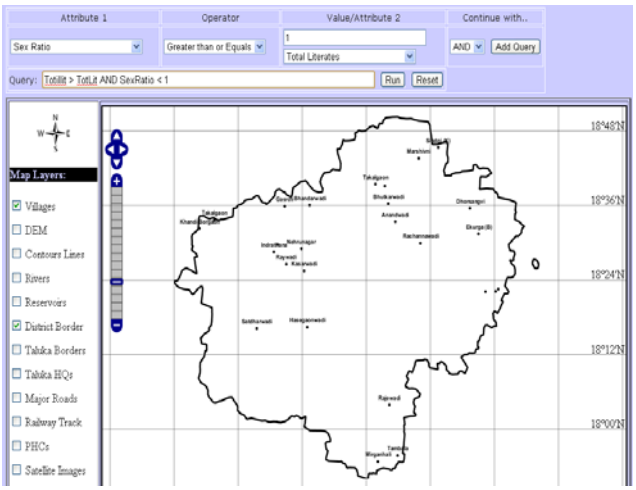


Figure 5. Filtered villages having more illiterates than literates and sex ratio less than 1.

Example 2:

Extract all villages in which their names either starts with character ‘S’ or ‘T’. Here the LIKE operator is used to complete this task. The following query is processed and the result is shown in Figure 6.

New_Map = SELECT * FROM VILLAGES1 WHERE Villages LIKE 'S%' OR Villages LIKE 'T%';

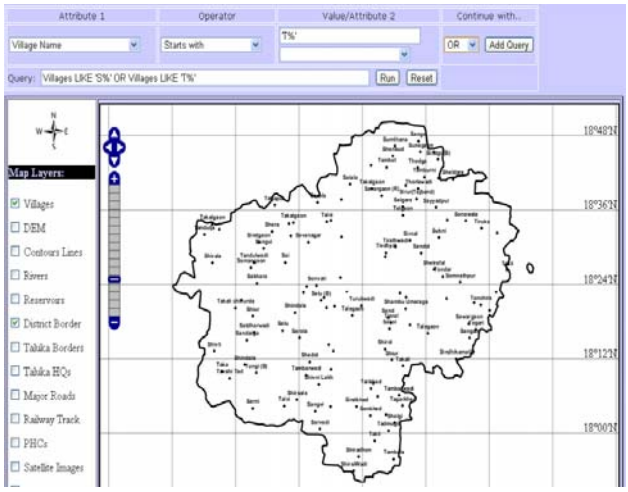


Figure 6. Villages' names starts with either 'S' or 'T' character.

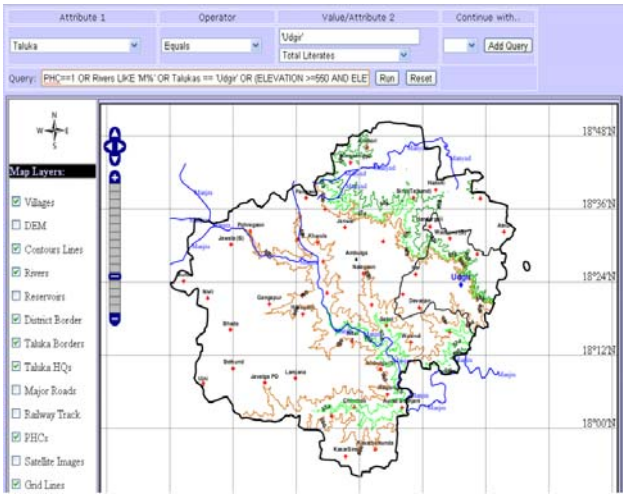


Figure 7. Extracted Geo-information from multiple attributes using multiple conditions.

Example 3:

The system can be used to extract/query information through clubbing more than one attributes at the same time. This example demonstrates that, one can extract village, elevation, river, taluka etc. information by executing a single query by clubbing multiple conditions of multiple attributes. The following are the few attributes used with different conditions within a single query to extract needful Geo-information.

- Show all villages, which are having PHCs.
- Show all rivers, whose name starts with 'M'
- Show taluka area and location, whose name is 'Udgir'
- Show all elevation contour poly lines, having greater than or equal to 550 AND less than or equal to 600 surface height.

The following query can be executed by selecting proper attribute 1, attribute 2 or user defined values and logical operators from filter attribute panel. The query will be

New_Map = PHC==1 OR Rivers LIKE 'M%' OR Talukas == 'Udgir' OR (ELEVATION >=550 AND ELEVATION<=600).

The Figure 7 shows the extracted information resulting from the above discussed query.

The main advantages of the system are:

- Centralized control over data & model resulting in lower costs for hardware, software, distribution, maintenance and training as well as more efficiency in model improvement and data update, particularly, for models with dynamic and real time information.
- The simple GUI and user friendly way to query/extract the village level information.
- Users do not need professional GIS knowledge, training or expensive and complex hardware and software as web based systems are platform independent.
- Allow public and stakeholders to access and participate in planning and decision making processes.
- Other facilities such as education, market changes, agriculture information and information of government schemes can be made available through the same system.

VII. Conclusion

In this paper, how one can design, develop and explore the spatial (geographical) information of a particular area, spread on internet in the form of web based spatial information system using open source geospatial tools. This chapter also deals with the following issues.

- Open source software can solve some of rural India's social, political, and administrative challenges and create a viable & cost effective technology for the provision of health, education, and other social services.
- Open source geospatial tools provide an added dimension to data analysis which helps in visualizing the real world complex problem.
- It's easy to modify using source codes of software tool rather than building new which facilitates the conservation of financial resources.
- Voluntary organizations can get better software tools for spatial and non spatial study helping in planning and management of development work.

REFERENCES

[1] G. Zeleke, B. Alemu, C. Hergarten and Juerg Krauer, Consultation Workshop on National Spatial Data Infrastructure and EthioGIS (2nd Release), University of Bern Switzerland, 2008, pp 69-81.

[2] A. Kumar, Participatory GIS for e-Governance and Local Planning: A Case Study in Different Environments on Its Applicability, 2008. Available: http://www.gisdevelopment.net/proceedings/mapworldforum/poster/MMW_Poster_52.pdf.

[3] G. Camara. (2003) Why Open Source GIS Software? Available: http://www.directionsmag.com/article.php?article_id=343&trv=1

[4] Espada P. G., Open for Change: Scoping paper on the use of FLOSS in Cadastre. and Land Registration Applications, FAO-NRLA, OSGeo, 2008

[5] GeoServer, 2008, Geographical Information System Web Server, website [Online]. Available: http://www.geoserver.org

[6] GMT, (2008), (Generic Mapping Tools), website. [Online]. Available: http://gmt.soest.hawaii.edu/gmt/gmt_home.html.

[7] Google Earth, 2009, Website [Online] Available: http://www.earth.google.com

[8] gvSIG, 2008, (Generalidad Valencia Sistema de Informacion Geografica) website. [Online]. Available: http://www.gvsig.gva.es.

[9] ILWIS, (2008), (The Integrated Land and Water Information System), website. [Online]. Available: www.itc.ni/ilwis/default.asp.

[10] Intermap's Product Handbook and Quick Start Guide, Version 3.3, 2004. http://www.intermap.com/images/handbook/producthandbook.pdf, Version 3.3.

[11] http://www.latur.nic.in/html/distprofile.html

[12] NRDB, Natural Resource Data Base, 2010, website [Online], Available: http://www.nrdb.co.uk

[13] OpenJUMP,(2008) (JAVA Unified Mapping Platform), website. [Online]. Available: http://openjump.org/, accessed on 20/10/2008.

[14] PostGIS, 2008, Suppoting object for PostGRESQL, website [Online]. Available: http://postgis.refractions.net

[15] PostGRESQL, 2008, ORDBMS Sql server for spatial data, website [Online]. Available: http://www.postgresql.org

[16] QGIS, (2008) website. [Online]. Available: www.qgis.org.

[17] Robinson, S., Turnbull, J., Brandner, S. and Fyfe, S., Can remote sensing be used to map vegetation and monitor community change in Antarctica?, Australian Antarctic Data Centre - CAASM Metadata., 2006 http://aadcmaps.aad.gov.au/aadc/metadata/metadata_edirect.cfm?md=AMD/AU/ASAC_2

[18] Seamless Data Server, 2008, website [Online], Available: www.seamless.usgs.gov/Website/Seamless/viewer.htm

[19] TerraView, (2008), website. [Online]. Available: http://www.terralib.org

[20] uDIG, (2008) (user-friendly desktop internet GIS), website. [Online]. Available: www.udig.refract

[21] Shalini Batra, Seema Bawa, Review of Machine Learning Approaches to Semantic Web Services Discovery, Journal of Advances in Information Technology, Vol. 1, No. 3, 2010, pp. 146-151.

Mr. Kodge Bheemashankar G. is working as a lecturer in department of studies and research in Computer Science of Swami Vivekanand College, Udgir Dist. Latur (MH) INDIA. He obtained MCM (Master in Computer Management) in 2004, M. Phil. in Computer Science in 2007 and registered for Ph.D. in 2008. His research areas of interests are GIS and remote sensing, digital image processing, data mining and data warehousing. He is published 34 research papers in national/international journals and conference proceedings.

Dr. P.S. Hiremath is a Professor and Chairman, Department of P. G. Studies and Research in Computer Science, Gulbarga University, Gulbarga-585106 INDIA, He has obtained M.Sc. degree in 1973 and Ph.D. degree in 1978 in Applied Mathematics from Karnataka University, Dharwad. He had been in the Faculty of Mathematics and Computer Science of Various Institutions in India, namely, National Institute of Technology, Surathkal (1977-79), Coimbatore Institute of Technology, Coimbatore(1979-80), National Institute of Technology, Tiruchirapalli (1980-86), Karnatak University, Dharwad (1986-1993) and has been presently working as Professor of Computer Science in Gulbarga University, Gulbarga (1993 onwards). His research areas of interest are Computational Fluid Dynamics, Optimization Techniques, Image Processing and Pattern Recognition. He has published 150+ research papers in peer reviewed International Journals and proceedings of conferences.

Design of Primary Screening Tool for Early Detection of Breast Cancer

Dr. C. Naga Raju
Professor and Head of IT, L.B.R.College of Engineering/Dept of IT, Mylavaram, India
Email: cnrcse@yahoo.com

C. Harikiran
Asst. Professor.
L.B.R.College of Engineering/Dept of IT, Mylavaram, India
Email: hari.ck10@gmail.com

T. Siva Priya
Asst. Professor.
L.B.R.College of Engineering/Dept of IT, Mylavaram, India
Email: sivapriya_tummala@yahoo.co.in

Abstract—The innovative approach consists of using the same algorithmic core for processing images to detect both microcalcifications and masses. Despite the advancement in the medical sciences cancer is claiming more than 50% of the people afflicted by it every year. Of all cancer incidence women around the world, the most commonly diagnosed type of non-skin cancer which results in death is Breast Cancer and this can be best detected by digital mammography. This paper includes the design and development of software expert system for real time mammogram image analysis. The system so designed would give the radiologist an idea about the exact shape and size of any tumor present in the breast. Radiologists however are unable to detect the cancerous growth when benign though it is detected in the mammograms due to varying criteria like dense flesh around the cancer or distractions due to neighboring features. This problem will be resolved by using Digital Image Processing techniques like Image Segmentation where the image will be segmented into similar regions by meaningfully assigning class labels to similar pixels in a region. Hence the cancerous growth will be detected in its early stage and the Radiologists will be able to do better diagnosis because Image Segmentation techniques are simple yet very effective. In this paper an innovative method is applied which consist mainly three steps. In the first step normalizes the regions in the breast images through uniform distribution of histogram equalization. In the second step fuzzy logic is applied to remove ambiguity in the misclassification region and in the third step a new Weight is applied to the previously extended OTSU method.

***Index Terms*—Breast Cancer, Mammogram, ROI, OTSU Threshold, Histogram Weight, Mean, Variance**

I. INTRODUCTION

Mammography and finding of suspicious masses during self-examinations form the primary screening tools for early detection of breast cancer [1, 2]. Early detection is difficult since the shape and size of the microcalcification clusters and speculated or irregular masses vary, and they are embedded in and camouflaged by various tissue structures [3, 4]. Mammography is the most effective method for the early detection of breast diseases. However, the typical diagnostic signs such as microcalcifications and masses are difficult to detect because mammograms are low-contrast images [5, 6].Breast cancer is considered as one of the primary causes of women mortality. The mortality rate in asymptotic women can be brought down with the aid of premature diagnosis. Despite the increasing number of cancers being diagnosed, the death rate has been reduced remarkably in the past decade due to the screening programs. Premature detection of breast cancer increases the prospect of survival whereas delayed diagnosis frequently confronts the patient to an unrecoverable stage and results in death [7, 8]. The Indian metropolises of Mumbai, Calcutta, and Bangalore display 23% of all the female cancers as breast cancers followed by cervical cancers [9, 10]. Despite the fact that the incidence of breast cancer in India is comparatively lower than that of the western countries, the issue is highly alarming. High quality images and mammographic interpretation are mandatory for the detection of premature and delicate symptoms of breast cancer. Mammogram (breast X-ray) is the medical image essential for the diagnosis of breast cancer and is considered to be the most dependable technique for premature detection. The widely recognized tool for the early detection of breast cancer in women

doi:10.4304/jait.3.4.228-235

with no symptoms; and to detect and diagnose breast disease in women experiencing symptoms like a lump, pain or nipple discharge, is mammography[11,12].

Contemporarily, screening mammography and radiographic imaging of the breast are the most effective tools for premature detection of breast cancer. Screening mammographic assessments are carried out on asymptomatic woman to detect premature, clinically unsuspected breast cancer. Still, studies have proved that all breast cancers that are retrospectively detected on the mammograms are not detected by radiologists. Due to the subtle and complex nature of the radiographic findings related with breast cancer, human factors such as varying decision criteria, distraction by other image features, and simple oversight can be responsible for the errors in radiological diagnosis. Computer assisted schemes that work on image processing and pattern recognition techniques can be utilized to enhance the diagnostic efficiency and for the location and classification of probable lesions and thereby alerting the radiologist to observe these areas with specific attention [13, 14]. Radiologists look out for particular abnormalities on mammograms visually. Some significant signs that radiologists pay attention to are clusters of microcalcifications, masses, and architectural deformations. A space-occupying lesion that is visible at more than one projection is referred to as a mass. Masses are illustrated with the aid of shape and margin features. Tiny deposits of calcium those are visible as minute bright spots on the mammogram are called as calcifications. They are exemplified by their type and distribution characteristics. The existence of microcalcifications is one of the significant and probably the only indication of cancer on a mammogram. A majority of the researches on computer analysis of mammograms have focused on the detection of small abnormalities, precisely the micro calcifications.

II. OTSU METHOD

Otsu thresholding proposed a criterion for maximizing the between-class variance of pixel intensity to perform picture thresholding [15, 16]. Basic OTSU Thresholding technique involves segmenting or decomposing the entire image into regions of some similar properties like pixels of same intensities for further analysis. Hence using this method the image can be separated into dark and light regions. This is called as Thresholding the image. The separated regions are called assigned class labels where the intensity levels of each pixel in one region will be greater than the Threshold value and the intensity levels of pixels in the second region will be less than the Threshold value. The high frequency components in the resultant image are enriched whereas the low frequency background structure was removed. A global threshold value applied on the reconstructed image acquired for each mammogram and a binary image providing all the probable points of microcalcifications formed. The threshold value is automatically obtained from the grey level histogram with the application of a peak detection method

A significant technique for image segmentation that attempts to recognize and extract a target from its background with the aid of the distribution of gray levels or texture in image objects is referred to as Thresholding [17]. The statistics of the one-dimensional (1D) histogram of gray levels and on the two-dimensional (2D) co-occurrence matrix of an image form the basis of a majority of the thresholding techniques. Precisely, the discriminant criterion chooses the optimal threshold in order to maximize the separability of the resultant classes in gray levels. The procedure makes use of only the zeroth- and the first-order cumulative moments of the gray-level histogram and hence is trouble-free[18]. It is possible to extend the method to multithreshold problems in an uncomplicated manner.

A. Methodology

An image is a 2D grayscale intensity function, and contains pixels with gray levels from 1 to L.
The probability of gray level in an image is:

$$P_i = f_i/N \Rightarrow \text{(number of pixels with gray level/total number of pixels)}$$

In the case of bi-level thresholding of an image; the pixels are divided into two classes, C_1 with gray levels [1, 2….t] and C_2 with gray levels [t+1….L].

Then, the gray level probability distributions for the two classes are

$$C_1 = p_1/w_1(t),\ldots.p_t/w_1(t)$$
$$C_2 = p_{t+1}/w_2(t),\ p_{t+2}/w_2(t),\ldots,p_L/w_2(t)$$

$$\text{Where } w_1(t) = \sum p_i \text{ (where i = 1, 2, 3..., t)}$$
$$\text{and } w_2(t) = \sum p_i \text{ (where i = t+1, t+2...L)}$$

Also, the means for classes and are

$$\mu_1 = \sum ip_i/w_1(t) \text{ (where i = 1, 2, 3..., t) and}$$
$$\mu_2 = \sum ip_i/w_2(t) \text{ (where i = t+1, t+2... L)}$$

Let μ_T be the mean intensity for the whole image. It is easy to show that

$$w_1\mu_1 + w_2\mu_2 = \mu_T \text{ and also } w_1 + w_2 = 1$$

Otsu defined the between-class variance of the threshold image as

$$\sigma_B^2 = w_1(\mu_1 - \mu_T)^2 + w_2(\mu_2 - \mu_T)^2$$

Likewise the above formula can be extended for use in case of multiple thresholds extension and Proposed method for basic OTSU method.

III. EXTENSION OF OTSU METHOD

The above OTSU method is simple and easier. However it fails if the Histogram is unimodal or close to unimodal. Hence an extension to the basic OTSU method will be implemented by selecting an optimal threshold. In this extended method the gray level distribution will be described using the average variance instead of average mean which is normally used in the basic OTSU method.

Here $\mu_1(t)$ and $\mu_2(t)$ can be regarded as the objects center gray and the background's center gray respectively, μ_T is the whole image center. This method makes sure that $(\mu_1 - \mu_2)^2$ is as bigger as it can get and gray distribution can be described not only by gray mean, but also by gray variance. The average variance will be used here to replace average mean in the basic OTSU method. The image variance reflects image uniformity; the variance is small inside of the objects and background. But the variance of edge and its neighborhood changes acutely. Hence it is reasonable to use average variance instead of the foreground and the background means in OTSU method.

$$t^* = \text{Arg Max}[w_1(\sigma_1^2(t) - \sigma_T^2(t))^2 + w_2(\sigma_2^2(t) - \sigma_T^2(t))^2$$

$$\sigma_0^2(t) = 1/w_0(t) \sum (i - \mu_1(t))^2 p(i) \text{ (where } i = 1, 2, 3..., t)$$
$$\sigma_1^2(t) = 1/w_1(t) \sum (i - \mu_1(t))^2 p(i) \text{ (where } i = t+1,.., m-1)$$
$$\sigma^2(t) = \sum (i - \mu_T(t))^2 p(i) \text{ (where } i = t+1, t+2, ..., L)$$

This method represents well adaptability and certain anti-noise abilities; it will not be although this method has some difficulties processing images with unimodal distribution.

IV. NOVEL FUZZY OTSU METHOD

Step1: Features are based on the grey-level histograms from selected regions of the breast. The distances to the skin normalized from 0 to 100 (providing invariance to the size of the breast) are utilized in the construction of the regions. Histogram modeling techniques alter an image in order to ensure that the histogram is of the desired shape. This is beneficial for the elongation of low levels of mammograms with the narrow histograms. Histogram equalization is a conventional histogram modeling methodology. According to the information theory, the uniform distribution attains maximum entropy, which encloses the most information. Thus, the mammogram information needs to be maximized in order to redistribute the gray-levels and achieve at the most uniform histogram .The next is fuzzy logic which produce optimal threshold to avoid the fuzziness in the image and makes good regions regarding background and object.

Step2: Fuzzy set theory assigns a membership degree to all elements among the universe of discourse according to their potential to fit in some class. The membership degree can be expressed by a mathematical function $\mu_A(x_i)$ that assigns, to each element in the set, a membership degree between 0 and 1.Let be the universe (finite and not empty) of discourse and x_i an element of .A fuzzy set **A** in **X** is defined as

$$A = \{(x_i, \mu_A(x_i)) | x_i \in X \}$$

The S –function is used for modeling the membership degrees. This type of function is suitable to represent the set of bright pixels and is defined as

$$\mu_{AS}(x) = s(x, a, b, c)$$

$$= \begin{cases} 0, & x \le a \\ 2\left\{\frac{(x-a)}{(c-a)}\right\}^2, & a \le x \le b \\ 1 - 2\left\{\frac{(x-c)}{(c-a)}\right\}^2, & b \le x \le c \\ 1, & x \ge c \end{cases}$$

Where $b = (1/2)(a + c)$

The S–function show in the Fig1 can be controlled through parameters **a** and **c**. Parameter **b** is called the cross over point where $\mu_{AS}(b) = 0.5.$ .The higher the gray level of a pixel (closer to white), the higher membership value and vice versa.

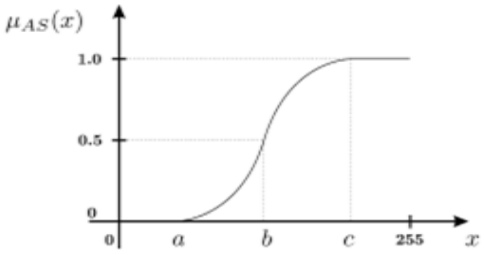


Fig1.Typical shape of the S-function function

Measures of Fuzziness are **a** reasonable approach to estimate the average ambiguity in fuzzy sets is measuring its fuzziness. The fuzziness of a crisp set should be zero, as there is no ambiguity about whether an element belongs to this t or not. If $\mu_A(x) = 0.5, \forall x,$,the set is maximally ambiguous and its fuzziness should be maximum. Degrees of membership near 0 or 1 indicate lower fuzziness, as the ambiguity decreases. Kaufmann in introduced an index of fuzziness (IF) comparing a fuzzy set with its nearest crispset. A fuzzy set **A*** is called crispset of **A** if the following conditions are satisfied:

$$\mu_{A^*}(x) = \begin{cases} 0, & if\ \mu_A(x) < 0.5 \\ 1, & if\ \mu_A(x) \ge 0.5 \end{cases}$$

Step3: The optimal threshold value exists at the valley of the two peaks or at the bottom rim of a single peak. The valley in the histogram that separates the object from

the background, its probability of occurrence is small in gray level histogram. Because of the optical threshold should near the cross where the object and the background intersect. The probability of occurrence at the threshold value should divide into two parts. Its first half belongs to background and second half belongs to object. Then we apply a new weight M to the OTSU method.

$$t = (P1*D1 + P2*D2) * M$$

$$\text{where } D1=(\sigma_1^2(o)- \sigma_T^2(t))^2$$

$$D2=(\sigma_2^2(b)- \sigma_T^2(t))^2$$

$$M=(1-P_T(t)/2)$$

Using this method we can make sure that the result threshold value resides at the valley or at the bottom of the right rim of single peak. It's also maximizes group variance and ensures that both the variance of the object and that of the background keep away from the variance of the whole image. Smaller the p(t)/2, larger will be the weight.

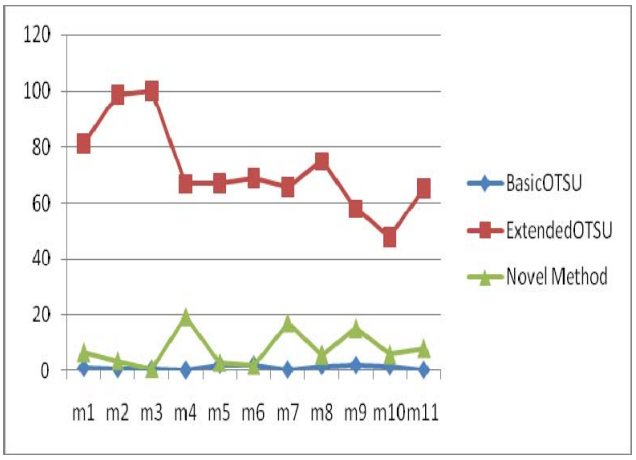
V. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the segmentation process using the proposed method, a set of images of different kinds were tested. Experimental results illustrate that the system is capable of aiding the interpretation of radiologists in their daily practice besides enhancing their diagnostic performance. The performance evaluation of three methods have been described based on table of values and graph shown in the paper .Results of grap1 reflect the severity of cancer using the above three methods. From the Basic OTSU method we can infer that the values plotted in graph1 are too low, hence we cannot clearly differentiate among the normal, moderate or severely cancer affected breast. The Extended OTSU method produced better results than basic OSTU method. But from the Extended OTSU method we can infer that the values plotted in graph1 are too high and ambiguous values among some of breast images, hence we cannot clearly differentiate among the normal, moderate or severely cancer affected breast. From the Novel Method we can infer that the values plotted in graph1 are neither too low nor too high (i.e., values are moderate). Further comparing the plotted values in graph1 against each of the images in table1 it is evident that the values clearly reflect the levels of severity of the cancer. For example the plotted values for images M4, M7, and M9 show that the severity of cancer in these respective images is too high, while the values for images M1, M8, M10, M11 show that the severity of cancer in these respective images is moderate and the values for images M2, M3, M5, M6 show that the severity of cancer in these respective images is low. Using the first two methods the levels of severity of cancer in respective images is not clearly reflected in the graph1. This is due to the fact that the values for Basic OTSU and Extended OTSU methods from table1 are either too low or too high. In the proposed method the defect can be extracted more precisely. It's able to select optimal threshold values for both unimodal and bimodal distribution, so it can be used on various defect detection mammograms applications.

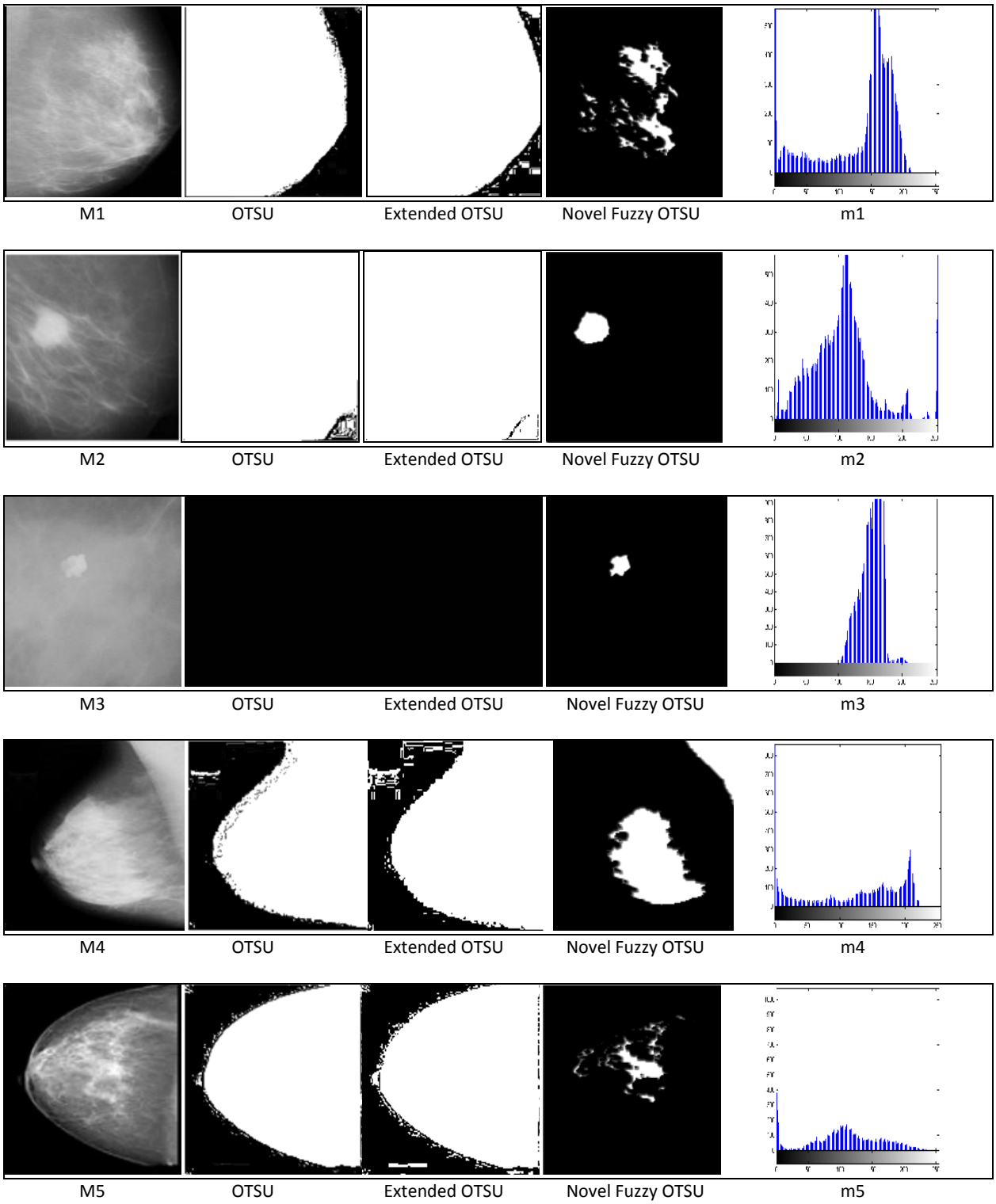
TABLE I.

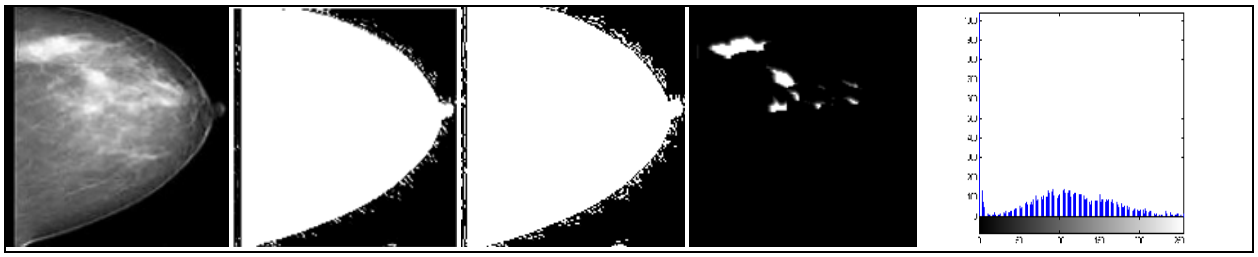
Images	BasicOTSU	ExtendedOTSU	Novel Method
M1	0.88	81.04	6.81
M2	0.47	98.65	3.71
M3	0.56	100	0.85
M4	0.13	66.57	19.17
M5	1.99	66.7	3.06
M6	2.03	68.49	2.04
M7	0.31	65.49	17.09
M8	1.69	74.67	5.84
M9	1.84	57.57	15.09
M10	1.47	47.45	6.04
M11	0.21	65.01	8.21

GRAPH1

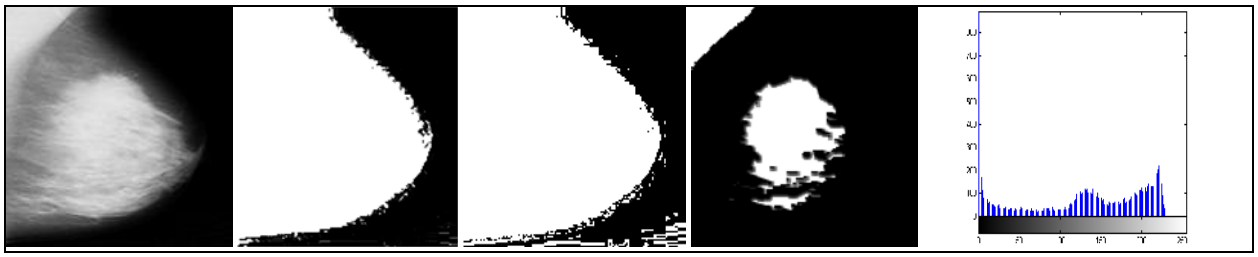


RESULTS

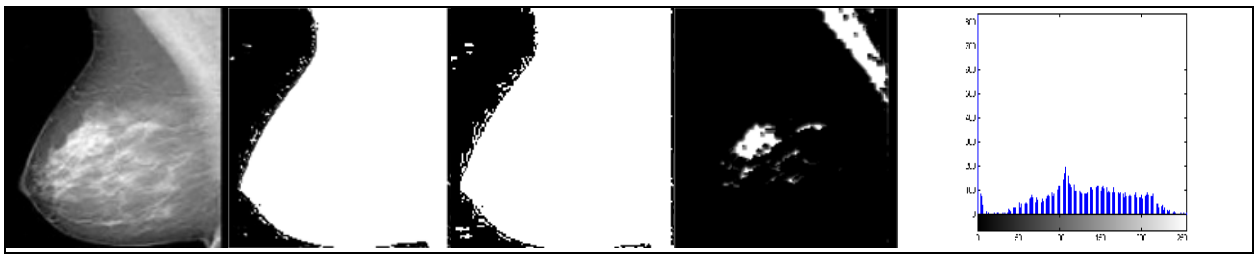




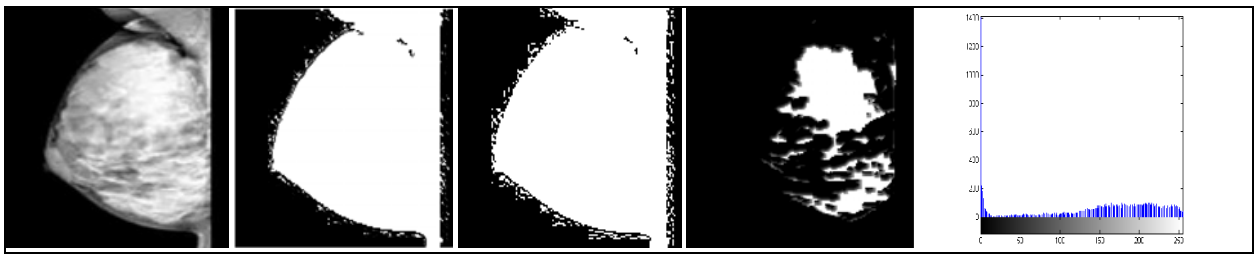
M6 OTSU Extended OTSU Novel Fuzzy OTSU m6



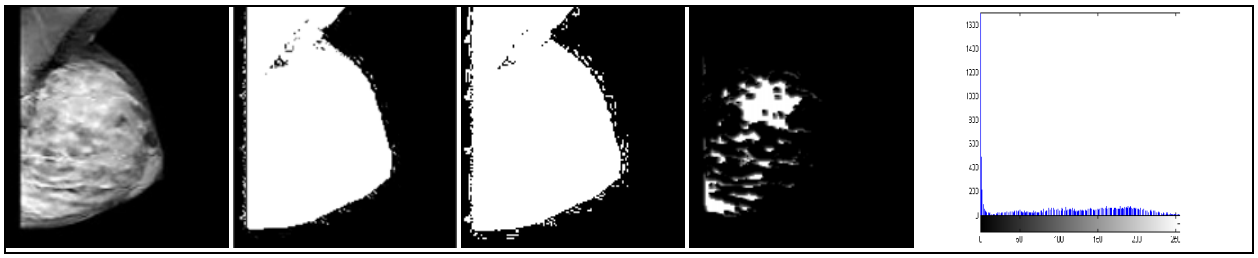
M7 OTSU Extended OTSU Novel Fuzzy OTSU m7



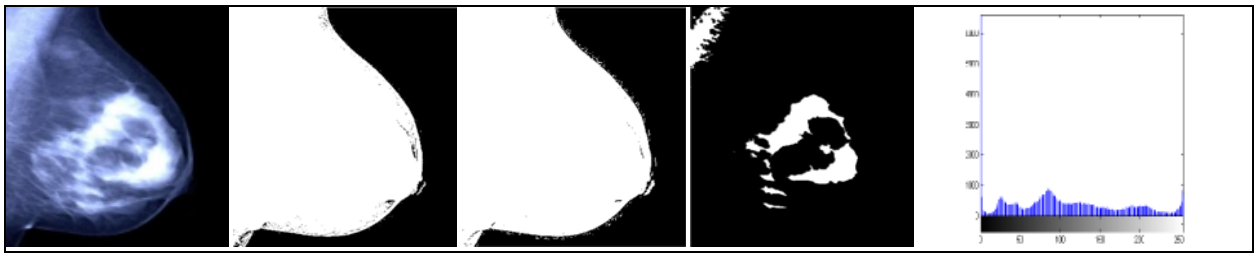
M8 OTSU Extended OTSU Novel Fuzzy OTSU m8



M9 OTSU Extended OTSU Novel Fuzzy OTSU m9



M10 OTSU Extended OTSU Novel Fuzzy OTSU m10



VI. CONCLUSION

The contemporary preference for the premature detection of breast cancer in women is Mammography. The elucidation of mammograms greatly depends on radiologist's opinion. The approach depends on an OTSU threshold operator strategy, for the segmentation of mass/microcalcification. In the proposed work we have assessed an automated detection method for one of the principal signs of breast cancer: clusters of microcalcifications and mass lesions. This technique involves normalizing regions of breasts and thresholding the Region of Interest (ROI) by using basic OTSU method. This method however fails if the Histogram is unimodal or close to unimodal. Hence an extension to the basic OTSU method will be implemented by selecting an optimal threshold. In this extended method the gray level distribution will be described using the average variance instead of average mean which is normally used in the basic OTSU method. So the average variance will replace average mean in the extended method. Furthermore extending this technique, a new method is proposed. In this method fuzzy logic is applied to remove ambiguity in the misclassification region and a new Weight is applied to the previously extended OTSU method. This proposed method ensures that both the variance of the object and variance of background are far from the variance of the image. This Weight ensures that the threshold is optimal and we will get satisfactory results for images with histogram of unimodal or multimodal distribution.

REFERENCES

[1] R.A. Smith, "Epidemiology of breast cancer in a categorical course in physics," Technical Aspects of Breast Imaging, 2nd ed., RSNA publication, Oak Book, II, pp.21, 1993.

[2] R. Peto, J. Boreham, M. Clarke, C. Davies., V. Beral, "UK and USA Breast cancer deaths down 25% in year 2000 at ages 20-69 years", THE LANCET, Volume 355, Issue 9217, Page 1822, 20 May 2000.

[3] Ranadhir Ghosh, Moumita Ghosh, John Yearwood, "A Modular Framework for Multi category feature selection in Digital mammography", In Proceedings of the 12th European Symposium On Artificial Neural Networks ESANN'2004, Bruges (Belgium), pp. 175-180, 28-30 April 2004.

[4] Bernard W. Stewart, Paul Kleihues, "WORLD CANCER REPORT", WHO, International Agency for Research on Cancer, IARC Press, Lyon 2003.

[5] E. L. Thurfjell, K. A. Lernevall, and A. A. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology, vol. 191, pp. 241–244, 1994.

[6] J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "Variability in radiologists' interpretations of mammograms," The New England journal of Medicine, vol. 331, no. 22, pp. 1493–1499, 1994.

[7] Huai Li, K. J. Ray Liu, and Shih-Chung B. Lo, "Fractal Modeling and Segmentation for the Enhancement of Microcalcifications in Digital Mammograms", IEEE Transactions on Medical Imaging, vol. 16, no. 6, December 1997.

[8] A. S. Pednekar and I. A. Kakadiaris, "Image segmentation based on fuzzy connectedness using dynamic weights," IEEE Trans. Image Process., vol. 15, no. 6, pp. 1555–1562, Jun. 2006.

[9] S. Sahaphong and N. Hiransakolwong, "Unsupervised image segmentation using automated fuzzy c-means," in Proc. IEEE Int. Conf. Computer and Information Technology, Oct. 2007, pp. 690–694.

[10] Veldkamp, W., Karssemeijer, N., "Accurate segmentation and contrast measurement of microcalcifications in mammograms: A phantom study", Medical Physics, vol. 25, pp. 1102-1110, 1998.

[11] L. Shen, R. Rangayyan, and J. Desaultels, "Detection and Classification Mammographic Calcifications", International Journal of Pattern Recognition and Artificial Intelligence. Singapore: World Scientific, pp. 1403–1416, 1994.

[12] K. Bowyer and S. Astley, "The Art of Digital Mammographic Image", Singapore: World Scientific, vol. 7, 1994.

[13] H. Barman, G. Granlund, and L. Haglund, "Feature extraction for computer- aided analysis of mammograms," in State of the Art of Digital Mammographic Image Analysis. Singapore: World Scientific, 1994, vol. 7, pp. 128–147.

[14] N. Otsu, A threshold selection method from gray-level histogram, IEEE Transactions on Systems Man Cybernet, SMC-8 pp. 62-66, 1978.

[15] H.F. Ng, "Automatic thresholding for defect detection", Pattern Recognition Letters, (27): 1644-1649, 2006.

[16] M. Sezgin and B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging, pp.146-156, 2003

[17] H. Lee and R. H. Park, Comments on an optimal threshold scheme for image segmentation, IEEE Trans. Syst.Man Cybern, SMC-20,741-742, 1990
[18] J. Z. Liu and W. Q. Li, The Automatic thresholding of gray-level picture via two-dimensional Otsu method, Acta Automatica Si.19, 101-105,1993

Dr. C. Naga Raju received his B.Tech degree in Computer Science from J.N.T.University Anantapur, M.Tech degree in Computer Science from J.N.T.University Hyderabad and PhD in digital Image processing from J.N.T.University Hyderabad. Currently, he is working as a professor & Head of IT in LakiReddy Bali reddy College of Engineering, Vijayawada. He is professor incharge for systems department. He has got 15 years of teaching experience. He has published thirty research papers in various national and international journals and about twenty eight research papers in various national and international conferences. He has attended twenty seminars and workshops. He is member of various professional societies like IEEE, ISTE and CSI.

Mr. C. Harikiran. He was born in Vijayawada, Andhra Pradesh, India. He received B.E (Bachelor of Engineering) degree in CSE, Computer Science & Engineering from University of Madras, Masters Degree M.Tech in Software Engineering Jawaharlal Nehru Technological University, Kakinada. He worked for over nine years as an IT Consultant in New York City. Presently he is working as Asst. Professor in Lakireddy Bali Reddy College of Engineering, Vijayawada. He plans to pursue PhD. His research interests are in the areas of Digital Image Processing, Artificial Intelligence, and Neural Networks, Information Security.

Siva Priya Tummala received her B.Tech degree in Information Technology from P.V.P S.I.T Vijayawada. She has got 2 years of teaching experience from Paladugu Parvathi Devi Institute of Engineering And Technology. Currently, she is working as Assistant Professor at Lakiredy Baliredy College of Engineering, Mylavaram. She has published two research papers in international journal.

Using Sequential Search Algorithm with Single level Discrete Wavelet Transform for Image Compression (SSA-W)

Mohammed Mustafa Siddeq
Software Engineering Dept.
Technical College – Kirkuk – IRAQ
Email: mamadmmx76@yahoo.com

***Abstract*— this research presents a new algorithm for an image compression consist of three phases; the first phase is using "Discrete Wavelet Transformation (DWT)", to produce low-frequency and high-frequencies sub-bands. The high-frequencies sub-bands are ignored (i.e. not used in this research), in the second phase used "Discrete Cosine Transformation (DCT)" applied on each "2x2" block from "LL" sub-band, then each block stored as a one-dimensional array in the new matrix called "Multi-Array-Matrix (MA-Matrix)". The third phase; MA-Matrix separated into "DC-Column" and "MA_2-Matrix", and then applied Minimize-Matrix-Size algorithm on the "MA_2-Matrix", to be as a one-dimensional array. Our decompression algorithm phase starts from "Sequential Search Algorithm (SS-Algorithm)" to find the estimated values for the "MA_2-Matrix". The SS-Algorithm depends on the three pointers, for decompress MA_2-Matrix, and then combined it with DC-Column for reconstructs MA-Matrix. Finally the inverse DCT and the inverse DWT are used for reconstructs approximately original image. Our approach compared with JPEG and JPEG2000 by using PSNR.**

***Index Terms*— Discrete Wavelet Transform, Discrete Cosine Transform, Minimize-Matrix-Size Algorithm, Sequential Search Algorithm.**

I. INTRODUCTION

Transform coding is at the heart of the majority of video coding systems and standards. Spatial image data (image samples or motion-compensated residual samples) are transformed into a different representation, the transform domain. There are good reasons for transforming image data in this way. Spatial image data is inherently ‘difficult’ to compress; neighboring samples are highly correlated (interrelated) and the energy tends to be evenly distributed across an image, making it difficult to discard data or reduce the precision of data without adversely affecting image quality. With a suitable choice of transform, the data is ‘easier’ to compress in the transform domain. There are several desirable properties of a transform for compression.

It should compact the energy in the image (concentrate the energy into a small number of significant values); it should de-correlate the data (so that discarding ‘insignificant’ data has a minimal effect on image quality); and it should be suitable for practical implementation in software and hardware [1,2].

The two most widely used image compression transforms are the discrete cosine transform (DCT) and the discrete wavelet transform (DWT) [3,4]. The DCT is usually applied to small, regular blocks of image samples (e.g. 8 x 8 squares) and the DWT is usually applied to larger image sections (‘tiles’) or to complete images. Many alternatives have been proposed, for example 3-D transforms (dealing with spatial and temporal correlation), variable block size transforms, and fractal transforms, Gabor analysis [5,7]. The DCT has proved particularly durable and is at the core of most of the current generation of image and video coding standards, including JPEG, H.261, H.263, H.263+, MPEG-l, MPEG-2 and MPEG-4. The DWT is gaining popularity because it can outperform the DCT for still image coding and so it is used in the new JPEG image coding standard (JPEG-2000) and for still ‘texture’ coding in MPEG-4[7,10].

This research introduces a proposed algorithm based on DWT, DCT and SS-Algorithm. The main idea for using two transformations is to increase number of zeros or insignificant coefficients, and make it easy to minimizing the size of "LL" sub-band into an array. In this research proposed new idea for matrix coding by using Minimize-Matrix-Size algorithm. Our approach compared with the techniques; JPEG and JPEG2000 by using PSNR.

II. PROPOSED COMPRESSION ALGORITHM (SSA-W)

Our image compression algorithm uses two frequency domains; 1) two-dimensional DWT, converts the images into four sub-bands, 2) two dimensional DCT applied on each "2x2" block of data from low-frequency sub-band "LL". These two transformations are used in our approach, for obtaining a lot of high-frequencies.

DCT plays main role for converting sub-band "LL" into DC-coefficients and AC-coefficients. DC-coefficients stored in the DC-column, while AC-coefficients are stored in the MA-Matrix, and then applied Minimize-Matrix-Size Algorithm on the MA-Matrix to convert it into an array, as shown in Figure-1.

doi:10.4304/jait.3.4.236-249

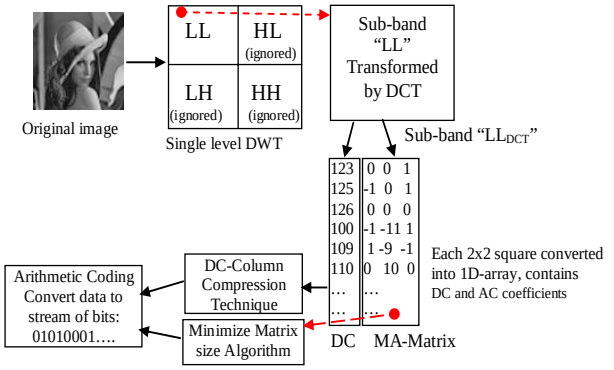


Figure 1. Proposed compression algorithm

A. Discrete Wavelet Transform (DWT)

The DWT it is the first transformation used in our algorithm, which is decompose an image into four frequencies bands: LL, LH, HL and HH. The LL is represented approximately an original image, and the other sub-bands; vertical, horizontal and diagonal are represented high frequency domain as shown in Figure-2 [1,3].

(a) Original image (b) Single Stage DWT

Figure 2. Image decomposed into 4 sub-bands

The wavelet transformation has some important properties; many of the coefficients for the high-frequency components (LH, HL and HH) are zero or insignificant [4,5,7]. This reflects the fact that the much of the important information in "LL" sub-band. The Daubechies wavelet transform has ability to reconstructs approximately original image just using low-frequency, while other sub-bands are ignored. These properties helped our algorithm to get higher compression ratio, but the decoded image has low quality.

B. Discrete Cosine Transform (DCT)

This is the second transformation used in this research, which is applies on each "2×2" block from "LL" sub-band as show in Figure-3.

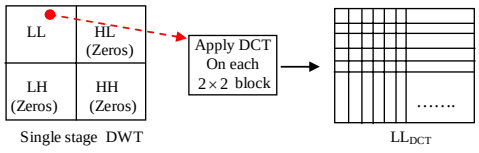


Figure 3. LL sub-band transformed again by DCT for each "2x2" block

The energy in the transformed coefficients is concentrated about the top-left corner of the matrix of coefficients. The top-left coefficients correspond to low frequencies: there is a 'peak' in energy in this area and the coefficients values rapidly decrease to the bottom right of the matrix, which means the higher-frequency coefficients [2,3]. The DCT coefficients are de-correlated, which means that many of the coefficients with small values can be discarded without significantly affecting in image quality. A compact matrix of de-correlated coefficients can be compressed much more efficiently than a matrix of highly correlated pixels. The following equations illustrated DCT and Inverse DCT function for two-dimensional matrices [7,9]:

$$C(u,v) = a(u)a(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (1)$$

$$where\ \backslash\quad a(u) = \sqrt{\frac{1}{N}}, \quad for u = 0$$

$$a(u) = \sqrt{\frac{2}{N}}, \quad for u \neq 0$$

$$f(x,y) = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1} a(u)\,a(v)\,C(u,v)\cos\left[\frac{(2x+1)\,u\pi}{2N}\right]\cos\left[\frac{(2y+1)\,v\pi}{2N}\right] \quad (2)$$

The difference between the applications of the Discrete Wavelet transformation (DWT) and Discrete Cosine transformation (DCT) is; the DWT typically apply to an image as one block or a large rectangular region of the image, while for DCT uses small block size [6,8,9]. Because the DCT becomes complicated to calculate for the larger block size, for this reason in this paper used just "2×2" block to be transformed by DCT, whereas a DWT will be more efficiently when applied on the complete image [10].

Each "2×2" coefficients from LL$_{DCT}$ divided by "QM", using matrix-dot-division and then truncate the result. This process called *Quantization Matrix*, which removes insignificant coefficients and increasing the zeros in LL$_{DCT}$. The QM can be computed as follows:

$$L = Quality \times \max(\text{Original Image}) \quad (3)$$

$$QM[i,j] = (i+j) + L \quad (4)$$

Note \ i , j =1,2

$$QM = \begin{array}{|c|c|}\hline 2+L & 3+L \\ \hline 3+L & 4+L \\ \hline \end{array}_{2\times2}$$

The parameter "L" in the above Eq(3) is computed from the maximum value in the original image (before transformation), and "*Quality*" value. The quality value is represents as an ratio for maximum value, if this ratio increased; the quality of an image is decreased

C. Minimize-Matrix-Size Algorithm

This section in our compression algorithm separating LL$_{DCT}$ into *DC-Column* and *MA$_2$-Matrix*, and then MA$_2$-

Matrix can be coded by using *Minimize-Matrix-Size* algorithm. This algorithm converts MA_2-Matrix into an array. Figure-4 shows LL_{DCT} separated into DC and MA_2.

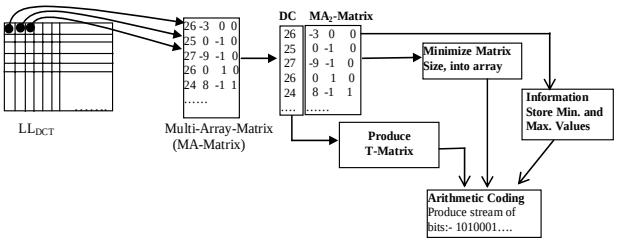


Figure 4. LL_{DCT} sub-band separated into DC-Column and MA_2-Matrix

Each " 2×2 " block from LL_{DCT} stored as a one-dimensional array in a new matrix called Multi-Array-Matrix (MA-Matrix). This matrix stores multi one-dimensional arrays, and the size of MA-matrix depends on the number of " 2×2 " blocks in the LL_{DCT} sub-band. Finally the MA-Matrix separated into DC-Column and MA_2-Matrix, then each one is compressing in a different way. The following equation represents the conversion MA_2-Matrix by using Random-Weights-Values into array:

$$DC(L)=MA(L,1) \quad (5)$$

$$Arr(L)=W(1)*MA_2(L,2)+W(2)*MA_2(L,3)+W(3)*MA_2(L,4) \quad (6)$$

Note\ "DC" represents DC-Column
"MA2" is represents second Multi-Array-Matrix
L=1,2,3,…size of MA-matrix

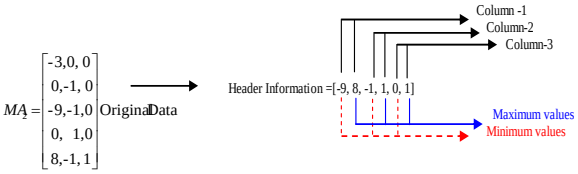
From the above Eq.(5), produced; "DC-Column" which is contains DC values, while "Arr" in Eq(6) contains floating point values. Now these two arrays are compressing by using arithmetic coding to produce stream of bits. The "W" in Eq.(6) are represents Random-Weights-Values are generated by using random function in MATLAB language, these values between {0…1}, for example: W={0.1, 0.65, 0.8519}.

The Random-Weights-Values are multiply with each row from MA_2-Matrix, to produce single floating point value "Arr", this idea it's similar to the Perceptron Neural Network equation. The neural networks; multiplies the weights values with input neurons to produce single output neuron [14,15]. List-1 illustrates Minimize-Matrix-Size Algorithm.

List -1 : Minimize-Matrix-Size Algorithm

```
Let Pos=1
W=Generate_Random_Weights_Values();
I=1;
While (I<row size LLDCT)
 J=1;
While (J<Column size LLDCT)
    %% immediately convert 2x2 blocks into array1x4
  L1x4=Convert_Block_into_Array( LLDCT[I, J])
  DC[Pos]=L1x4[1]   %%  store first value in DC-Column
  Arr[Pos]=0;       %% initialize before make summation
  For K=2 to 4
  Arr[Pos]= Arr[Pos] + L1x4[K] * W(K-1);
  End; %% End for
    Pos++;
 J=J+2;
End; %% End while
 I=I+2;
End;  %% End while
```

Before applying Minimize-Matrix-Size Algorithm, our compression algorithm stores minimum and maximum original values for MA_2-Matrix as a header-information, this is because the decompression algorithm SS-Algorithm depends on these original values to reconstruct MA_2-Matrix (See Section 3). The minimum and maximum values for each column are stored independently. The following example illustrates storing original minimum and maximum values as a header information will be used later by our decompression algorithm:-



D. Apply one-dimensional DCT on DC-Column (T-Matrix)

The DC-Column contains integer values, these values are obtained from previous section (See Section "*C*"). The arithmetic coding unable to compress these values, for this reason DC-Column partitioned into 128-arrays, each array transformed by one-dimensional DCT (See Eq.(1)) by changing u=0, x=0, then each array is quantized and stored as a row in the matrix called Transformed-Matrix (T-Matrix). The quantization process applied on each array, according to the following equation:

$$Q(n)=Q(n-1)+2 \quad (7)$$

Note\ *n=2,3,..128, the initial value of Q(1)=12.8*

The first value in Q(n) is 12.8, because the length of each row in T-Matrix is 128, as shown in Figure-5.

The idea for using T-Matrix, this is because we need to makes de-correlated values, for obtaining good compression ratio. Each row from T-Matrix consists of; low-frequency and high-frequency, now we need to scanning column-by-column for converting the T-Matrix into one-dimensional array.

Run Length Encoding (RLE) and Arithmetic Coding playing main role to compress the one-dimensional array as a lossless data compression. RLE it counts the repeated coefficients, and then refers to the repeated coefficients by a value, this process reduce the length of the repeated data, then applying the arithmetic coding to convert these reduced data into bits. For more information about RLE and Arithmetic coding is in the reference [4].

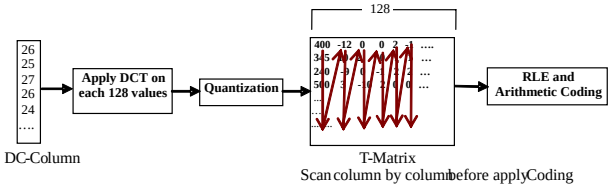


Figure 5. DC-Column compression technique

III. Decompression Algorithm by Sequential Search Algorithm (SS-Algorithm)

Our decompression algorithm starts with SS-Algorithm, which is designed to be inverse Minimize-Matrix-Size algorithm. SS-Algorithm it is new algorithm used for decode MA_2-Matrix, this algorithm depends on the coded "Arr" and Random-Weights-Values for reconstructs MA_2-Matrix.

Our decompression algorithm consists of three phases; the first phase; decodes DC-Column values (See Figure 5). The second phase; using SS-Algorithm for reconstructs the MA_2-Matrix, at the third phase, combines DC-Column with MA_2-Matrix to produce decoded MA-Matrix. The following figure shows the first phase of our decompression algorithm (i.e. decoding DC-Column).

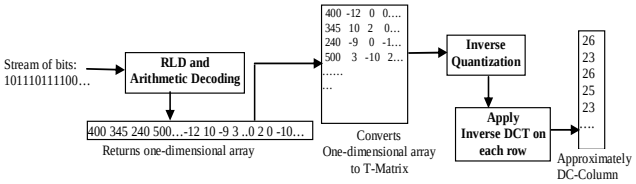


Figure 6. First phase for our decompression algorithm "Decoding DC-Column"

Run Length Decoding (RLD) and Arithmetic decoding, produced one-dimensional-array contains the original data for the T-Matrix. This array placed at columns of the T-Matrix. After generates the T-Matrix, and then applied *Inverse Quantization*, It is dot-multiplication of the row coefficients with "Q(n)" (See Eq.(7)), then one-dimensional inverse DCT applied on each row to produce approximately original 128-values (See Eq.(2)), this process will continues until all rows in T-Matrix are completes .

The second phase in our decompression algorithm it is very important for decoding MA_2-Matrix. In this phase uses SS-Algorithm which is depends on the coded "Arr" with header information (i.e. minimum and maximum values for MA_2-Matrix). The following figure shows second phase for our decompression algorithm depends on SS-Algorithm:

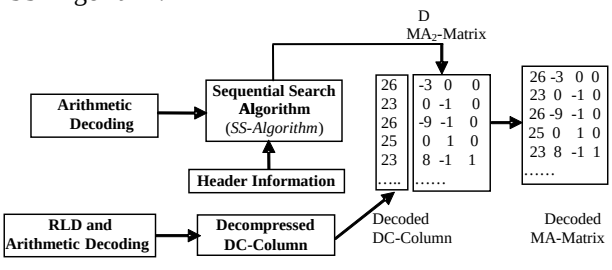


Figure 7. Second phase for our decompression algorithm "SS-Algorithm decoding steps"

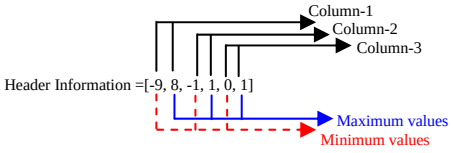
The SS-Algorithm used for searching missing data within minimum and maximum values. To explain in details, assume the following MA_2-Matrix contains 3*5 data:

$$MA_2 = \begin{bmatrix} -3, 0, 0 \\ 0, -1, 0 \\ -9, -1, 0 \\ 0, 1, 0 \\ 8, -1, 1 \end{bmatrix} \text{ Original Data}$$

$$Arr = [-0.3, -0.65, -1.55, 0.65, 1.0019]$$

Coded Data by "Minimize-Matrix-Size Algorithm"

The above "MA_2" data are compressed by using Minimized-Matrix-Size algorithm to produce "Arr" (See List-1), and header information about coded data is represents minimum and maximum values for each column in MA_2-Matrix, as shown below:



The SS-Algorithm decompression concepts start from three pointers; S1, S2 and S3, these pointers responsible for decoding column-1, column-2 and column-3 respectively. these pointers; S1, S2 and S3 are increasing by one at each iteration, as an digital clock, starting increment from S1, after reaching to the limited value (i.e. maximum value), S2 will start to increment, then S3 and this means the S1 is represents inner loop, while S2 and S3 represents outer loop respectively. The SS-Algorithm sets pointers to minimum values, taken from the "Header Information". Then at each iteration SS-Algorithm uses Eq(9) to compares the result of "Sum" with "Arr" (See Eq(6)) and the "Error" must be equal to zero, which means the pointers; S1, S2 and S3 are represents estimated values for each row in MA_2-Matrix.

$$Sum = \sum_{L=1}^{3} S(L) * W(L) \qquad (8)$$

Note\ S(L) :- represents S1, S2 and S3

$$Error = |Arr - Sum| \qquad (9)$$

For test SS-Algorithm, assume the "Arr" values computed from Minimized-Matrix-Size algorithm for the above example "MA_2". the algorithm start to find the first row values at "MA_2", depending on Arr(1)= -0.3, the initial values will be S1=-9, S2=-1 and S3=0 according to the header information, then "S1" begins to increment by one at each iteration, then "S2" begins to increment once, and then "S1" back to beginning (i.e. minimum value) for increment again. Finally if "S2" reached to maximum value "S3" will begins to increment, while "S1" and "S2" will back to beginning. This process will continue until finds the estimated "MA_2" values. SS-Algorithm at each increment the Eq.(9) will checked if it is zero or not. The

TABLE-1 shows the number of increments and number of iterations to find the estimated values for the example:

Arr = { - 0.3, - 0.65, -1.55, 0.65, 1.0019 }

TABLE I
SS-ALGORITHM ITERATIONS AND INCREMENTS TO ESTIMATE MA_2-MATRIX

Row for MA_2-Matrix	Number of increments			Number of Iterations	Estimated values		
	S1	S2	S3		S1	S2	S3
1st	25	1	0	25	-3	0	0
2nd	10	0	0	10	0	-1	0
3rd	0	0	0	0	-9	-1	0
4th	36	2	0	36	0	1	0
5th	54	2	1	54	8	-1	1

In the above TABLE I, the number of total addition for "S1" represents number of iterations, because "S1" increments 18 times, and then "S2" will work, and "S1" returns to the minimum value, then "S1" increments 7 times to reach to the estimated values, this means the total number of iteration is 25 iterations to find just 1st row in MA_2. List-2 illustrates SS-Algorithm:

List-2 Sequential Search Algorithm (SS-Algorithm**)**

```
%% minimum value for 1st  column
Let   S1_Min=Headr_information[1].min;
%% maximum value for 1st  column
Let   S1_Max= Headr_information[2].max;
%% minimum value for 2nd   column
Let   S2_Min= Headr_information[3].min;
%% maximum value for 2nd   column
Let   S2_Max= Headr_information[4].max;
%% minimum value for 3rd   column
Let   S3_Min= Headr_information[5].min;
%% maximum value for 3rd   column
Let   S3_Max= Headr_information[6].max;

For I=1 to size of Arr
        S1 = S1_Min; %% set pointers to initial values
        S2 = S2_Min;
        S3 = S3_Min;
Error =1;

    While (Error not equal to Zero )
        Sum = S1 * W(1)+ S2 * W(2)+ S3 * W(3);
          Error = | Sum – Arr[I] |
         IF (Error == Zero)
          Stop While  loop();
         End
          S1++;   %% increment by one
        IF (S1 > S1_Max) S1=S1_Min; S2++ End;
        IF (S2 > S2_Max) S2=S2_Min; S3++ End;
        IF (S3 > S3_Max) S3=S3_Min;        End;
     End; %% End while
     %% Decode  MA2 by using S1, S2, S3
    MA2[I][1]=S1;  MA2[I][2]=S2;  MA2[I][3]=S3;
End; %%  End for
```

After the SS-Algorithm estimated MA_2-Matrix, at the third phase MA_2-Matrix combined with DC-Column to reconstruct MA-Matrix, then each row in MA-Matrix is converted to "2 x 2" block in LL_{DCT}. Inverse quantization (See Eq.(4)) and inverse DCT (See Eq.(2)) applied on each block respectively to get the decoded LL sub-band. This leads for decoding approximately original image by using inverse DWT.

IV. COMPUTER RESULTS

In this section, our compression algorithm tested on the three gray level images "LENA", "X-ray" and "Girl". These images are tested by Microprocessor AMD Athalon - 2.1GHz with 3G RAM and using the MATLAB R2008a as a programming language with O.S. Windows 7(32bit). Our compression algorithm (SSA-W) uses; single stage Daubechies DWT (db3), then applies DCT on each "2x2" block from LL sub-band. Each "2x2" block are quantized by QM (See Eq. (3)), and then using Minimize-Matrix-Size algorithm to compress LL_{DCT}. TABLE II shows our compression results. The decompressed images; "LENA", "X-ray" and "Girl" are shows in Figure-8, Figure-9 and Figure-10.

TABLE II
SSA-W COMPRESSION TECHNIQUE APPLIED ON THE IMAGES

Image Name	*Parameter "Quality" used in QM (Kbytes)*	*Compressed Image size (Kbytes)*	*Information for compressed image (Kbytes)*	*Total Compressed Image size (Kbytes)*
Lena dimension (500x500) Original Image size 245 Kbytes	0.02	15.58 Kb	0.75 Kb	16.33 Kb
	0.05	10.52 Kb	0.64 Kb	11.16 Kb
	0.08	7.99 Kb	0.58 Kb	8.57 Kb
	0.1	6.8 Kb	0.54 Kb	7,34 Kb
	0.15	5 Kb	0.48 Kb	5.48 Kb
X-ray dimension (500x522) Original Image size 255 Kbytes	0.02	11.3 Kb	0.77 Kb	12.07 Kb
	0.05	7.6 Kb	0.67 Kb	8.27 Kb
	0.08	5.7 Kb	0.59 Kb	6.29 Kb
	0.1	4.8 Kb	0.55 Kb	5.35 Kb
	0.15	3.3 Kb	0.47 Kb	3.77 Kb
Girl dimension (595x774) Original Image size 451 Kbytes	0.02	33 Kb	1.4 Kb	34.4 Kb
	0.05	24.3 Kb	1.2 Kb	25.5 Kb
	0.08	19.4 Kb	1.1 Kb	20.5 Kb
	0.1	17.1 Kb	1 Kb	18.1 Kb
	0.15	13 Kb	0.92 Kb	13.92 Kb

(a) Original image dimension (500 x500)

(b) Decoded image with Quality =0.02, PSNR=33.5 dB

(e) Decoded image with Quality=0.1, PSNR= 32.2 dB

(c) Decoded image with Quality=0.05, PSNR=33.3 dB

(f) Decoded image with Quality =0.15, PSNR=31.5 dB

Figure 8. (a - f) Decompression "LENA" image by our compression approach (SSA-W) according to parameter "Quality"

(d) Decoded image with Quality =0.08, PSNR=32.8 dB

(a) Original image dimension (500 x522)

(b) Decoded image with Quality =0.02, PSNR=30 dB

(c) Decoded image with Quality=0.05, PSNR=29.9 dB

(d) Decoded image with Quality = 0.08, PSNR=29.9 dB

(e) Decoded image with Quality=0.1, PSNR=29.6 dB

(f) Decoded image with Quality = 0.15, PSNR=29.6 dB

Figure 9. (a - f) Decompression "X-ray" image by our compression approach (SSA-W) according to parameter "Quality"

(a) Original image dimension (595x 774)

(d) Decoded image with Quality = 0.08, PSNR=31.4 dB

(b) Decoded image with Quality =0.02, PSNR=31.9 dB

(e) Decoded image with Quality=0.1, PSNR=30.9 dB

(c) Decoded image with Quality=0.05, PSNR=31.6 dB

(f) Decoded image with Quality = 0.15, PSNR=30.6 dB

Figure 10. (a - f) Decompression "Girl" image by our compression approach (SSA-W) according to parameter "Quality"

TABLE III
RESULT FOR SSA-W DECOMPRESSION TECHNIQUE FOR "LENA" IMAGE

Parameter "Quality" used in QM	*Execution time for SS-Algorithm to find the decoded image*	*PSNR*
0.02	53.9 (Sec.)	33.5 dB
0.05	10.49 (Sec.)	33.3 dB
0.08	5.75 (Sec.)	32.8 dB
0.1	4.1 (Sec.)	32.2 dB
0.15	2 (Sec.)	31.5 dB

TABLE IV
RESULTS FOR SSA-W DECOMPRESSION TECHNIQUE FOR "X-RAY" IMAGE

Parameter "Quality" used in QM	*Execution time for SS-Algorithm to find the decoded image*	*PSNR*
0.02	151 (Sec.)	30 dB
0.05	18.2 (Sec.)	29.9 dB
0.08	6.9 (Sec.)	29.9 dB
0.1	5.4 (Sec.)	29.6 dB
0.15	2.3 (Sec.)	29.6 dB

TABLE V
RESULTS FOR SSA-W DECOMPRESSION TECHNIQUE FOR "GIRL" IMAGE

Parameter "Quality" used in QM	*Execution time for SS-Algorithm to find the decoded image*	*PSNR*
0.02	333 (Sec.)	31.9 dB
0.05	34.4 (Sec.)	31.6 dB
0.08	15.7 (Sec.)	31.4 dB
0.1	10.9 (Sec.)	30.9 dB
0.15	7.8 (Sec.)	30.6 dB

Peak Signal to Noise Ratio (PSNR) its used in TABLE III, TABLE IV and TABLE V, refers to the image quality mathematically, PSNR it is a very popular quality measure, can be calculates very easily between The decompressed image and the original image, as shown in the following equation[1-3]:

$$\mathrm{PSNR} = 10\log_{10}\frac{255^2}{\mathrm{MSE}} \tag{10}$$

$$MSE = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|Decode(i,j) - Original(i,j)\right|^2 \tag{11}$$

Our compression algorithm SSA-W is compared with JPEG and JPEG2000; these two techniques are used widely in the digital image compression, especially for image transmission and video compression. The JPEG technique based on the DCT applied on the partitioned image, then each partition are encoded by RLE and Huffman encoding. For more information about it read the references [4,11]. The JPEG2000 based on the multi decomposition for DWT applied on the also partitioned image and then each partition encoded by Arithmetic encoding [12,13]. The new ACDSee application has many options for save the images as JPEG and JPEG2000 formats. The most options are used is the "Quality". If the image "Quality" is increased, the compression ratio will be decreased and vice versa. TABLE VI, TABLE VII, TABLE VIII shows the results of the compressed images; "Lena", "X-ray" and "Girl" respectively, by the two techniques and compared with our compression approach according to PSNR.

TABLE VI
COMPARISON BETWEEN JPEG, JPEG2000 AND SSA-Q FOR "LENA" IMAGE

Method	*Quality*	*Compressed image size*	*PSNR*
JPEG	26 %	16.5 Kbytes	33 dB
	12%	11 Kbytes	30.4 dB
	7%	8.46 Kbytes	28.4 dB
	5 %	7.33 Kbytes	26.9 dB
	1 %	5.87 Kbytes	23.9 dB
JPEG2000	76%	16.5 Kbytes	35.6 dB
	67%	11 Kbytes	33.8 dB
	60%	8.5 Kbytes	32.7 dB
	56 %	7.49 Kbytes	32.2 dB
	45 %	5.46 Kbytes	30.9 dB
Proposed SSA-W	0.02	16.33 Kbytes	33.5 dB
	0.05	11.16 Kbytes	33.3 dB
	0.08	8.57 Kbytes	32.8 dB
	0.1	7.34 Kbytes	32.2 dB
	0.15	5.48 Kbytes	31.5 dB

TABLE VII
COMPARISON BETWEEN JPEG, JPEG2000 AND SSA-W FOR "X-RAY" IMAGE

Method	*Quality*	*Compressed image size*	*PSNR*
JPEG	27 %	12 Kbytes	31.8 dB
	15%	8.6 Kbytes	30.1 dB
	8%	6.3 Kbytes	28.2 dB
	5 %	5.29 Kbytes	26.4 dB
	1 %	4.4 Kbytes	24.1 dB
JPEG2000	69%	12.5 Kbytes	34.8 dB
	58%	8.36 Kbytes	33.3 dB
	49%	6.3 Kbytes	32.3 dB
	43 %	5.25 Kbytes	31.6 dB
	28 %	3.87 Kbytes	30.3 dB
Proposed SSA-W	0.02	12.07 Kbytes	30 dB
	0.05	8.27 Kbytes	29.9 dB
	0.08	6.29 Kbytes	29.9 dB
	0.1	5.35 Kbytes	29.6 dB
	0.15	3.77 Kbytes	29.6 dB

TABLE VIII
COMPARISON BETWEEN JPEG, JPEG2000 AND SSA-W FOR "GIRL" IMAGE

Method	*Quality*	*Compressed image size*	*PSNR*
JPEG	23 %	33.7 Kbytes	31.4 dB
	14%	25.2 Kbytes	29.5 dB
	10%	20.6 Kbytes	28.2 dB
	8%	18 Kbytes	27.2 dB
	6 %	15.1 Kbytes	26 dB
JPEG2000	79%	35 Kbytes	34.7 dB
	72%	24.8 Kbytes	32.4 dB
	68%	20.8 Kbytes	31.3 dB
	64%	17.8 Kbytes	30.5 dB
	57%	13.8 Kbytes	29 dB
Proposed SSA-W	0.02	34.4 Kbytes	31.9 dB
	0.05	25.5 Kbytes	31.6 dB
	0.08	20.5 Kbytes	31.4 dB
	0.1	18.1 Kbytes	30.9 dB
	0.15	13.92 Kbytes	30.6 dB

The PSNR it's not represents a perfect measurement for image quality, this is because the MSE computes the square for the differences, for example; the MSE between "10" and "14" is "16". This means the difference is big between "10 and "14", actually the difference between these two numbers are small according to the brightness.

For this reason we need for the Human Visual System as a second measurement for comparison between decompressed images by JPEG, JPEG2000 and SSA-W to show the details for the decoded images [4]. This research shows part from each decoded image and zoomed-in one time, to show the features for each method. Figure-11, Figure-12 and Figure-13 shows the comparison for the decoded images; "Lena", "X-ray" and "Girl" respectively.

(a) Decoded by **SSA-W**, Quality=0.02

(b) Decode by **JPEG2000**, Quality = 76%

(c) Decode by **JPEG**, Quality = 26%

(d) Decoded by **SSA-W**, Quality=0.05

(e) Decode by **JPEG2000**, Quality = 67%

(f) Decode by **JPEG**, Quality = 12%

(g) Decoded by **SSA-W**, Quality=0.08

(h) Decode by **JPEG2000**, Quality = 60%

(i) Decode by **JPEG**, Quality = 7%

(j) Decoded by **SSA-W**, Quality=0.1

(k) Decode by **JPEG2000**, Quality = 56%

(l) Decode by **JPEG**, Quality = 5%

(m) Decoded by **SSA-W**, Quality=0.15 (n) Decode by **JPEG2000**, Quality = 45% (o) Decode by **JPEG**, Quality = 1%

Figure 11. (a - o) Comparison between SSA-W, JPEG2000 and JPEG by using part of zoomed-in "LENA" image

(a) Decoded by **SSA-W**, Quality=0.02 (b) Decoded by **JPEG2000**, Quality=69% (c) Decoded by **JPEG**, Quality=27%

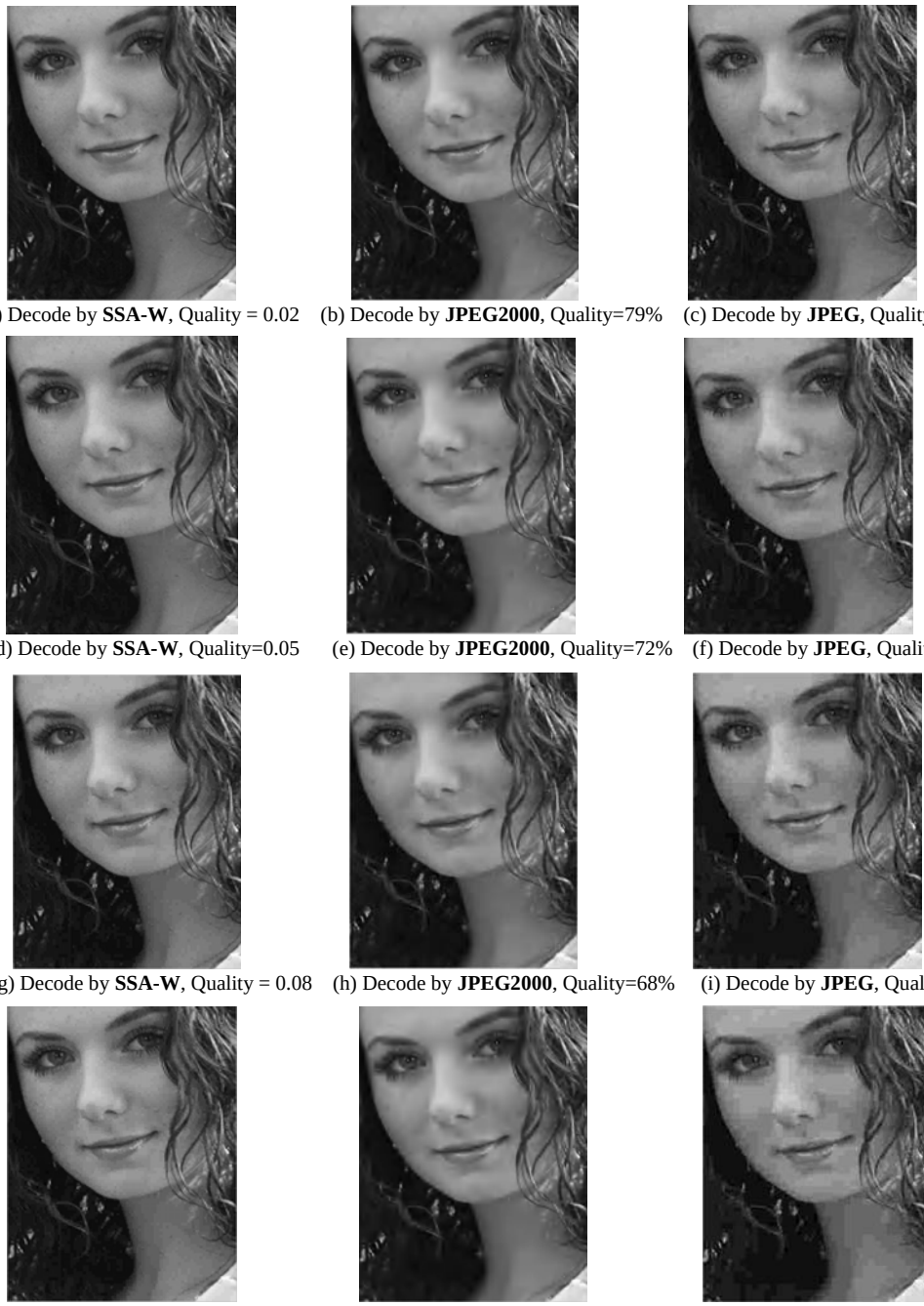
(d) Decoded **by SSA-W**, Quality=0.05 (e) Decoded by **JPEG2000**, Quality=58% (f) Decoded by **JPEG**, Quality=15%

(g) Decoded **by SSA-W**, Quality=0.08 (h) Decoded by **JPEG2000**, Quality=49% (i) Decoded by **JPEG**, Quality=8%

(j) Decoded by **SSA-W**, Quality=0.1 (k) Decoded by **JPEG2000**, Quality=43% (l) Decoded by **JPEG**, Quality=5%

(m)Decoded by **SSA-W**, Quality=0.15 (n) Decoded by **JPEG2000**, Quality=28% (o) Decoded by **JPEG**, Quality=1%

Figure 12. (a - o) Comparison between SSA-W, JPEG2000 and JPEG by using part of zoomed-in "X-ray" image



(a) Decode by **SSA-W**, Quality = 0.02 (b) Decode by **JPEG2000**, Quality=79% (c) Decode by **JPEG**, Quality=23%

(d) Decode by **SSA-W**, Quality=0.05 (e) Decode by **JPEG2000**, Quality=72% (f) Decode by **JPEG**, Quality=14%

(g) Decode by **SSA-W**, Quality = 0.08 (h) Decode by **JPEG2000**, Quality=68% (i) Decode by **JPEG**, Quality=10%

(j) Decode by **SSA-W**, Quality = 0.1 (k) Decode by **JPEG2000**, Quality=64% (l) Decode by **JPEG**, Quality=8%


(m) Decode by **SSA-W**, Quality=0.15


(n) Decode by **JPEG2000**, Quality=57%


(o) Decode by **JPEG**, Quality=6%

Figure 13. (a - o) Comparison between SSA-W, JPEG2000 and JPEG by using part of zoomed-in "Girl" image

V. CONCLUSIONS

This research introduces a proposed algorithm for image compression, based on the two important transformations for coding and SS-Algorithm for decoding. This research has some advantages illustrated in the following steps:

1- Using two transformations, helped our compression algorithm for increasing number of high-frequency coefficients, and leads to increases compression ratio.
2- Minimize Matrix Size algorithm is used to collect three coefficients from MA_2-Matrix, to be single floating point value. This process converts the matrix into array, and this leads to increasing compression ratio, and in another hand keeps the quality of the high-frequency coefficients.
3- The properties of the Daubechies DWT (db3) helps our approach to obtain higher compression ratio, this is because the high-frequencies are ignored in our approaches. This is because the Daubechies DWT family has the ability to zooming-in an images, and not necessarily needs for the high-frequencies sub-bands.
4- SS-Algorithm is represents the core of our decompression algorithm, converts the one-dimensional array into the MA_2-Matrix. This algorithm is depending on the Random-Weights-Values.
5- SS-Algorithm finds solutions as faster as possible with three pointers.
6- Our approach gives good visual image quality, more than JPEG and JPEG2000. This is because our approach removes the blurring that caused by Multi-level DWT in JPEG2000 and removes most of the block artifacts that caused by 8×8 DCT in JPEG .

Also this research has some disadvantages illustrated in the following steps:

1- The differences between JPEG2000 and SSA-W are; JPEG2000 faster than SSA-W when obtains high image quality, while our approach SSA-W takes more execution time for obtains high quality images.
2. The main reason for makes SS-Algorithm slower; the range between minimum and maximum values. This makes our approach slower than JPEG and JPEG2000 for obtains high quality images.

REFERENCES

[1] G.Sadashivappa and K.V.S.Ananda Babu, "PERFORMANCE ANALYSIS OF IMAGE CODING USINGWAVELETS", IJCSNS *International Journal of Computer Science and Network Security*, VOL.8 No.10, 2008.
[2] Ali Al-Haj, "Combined DWT-DCT Digital Image Watermarking", *Science Publications, Journal of Computer Science* 3 (9): 740-746, 2007.
[3] Grigorios D. , N. D. Zervas, N. Sklavos and Costas E. Goutis "Design Techniques and Implementation of Low Power High-Throughput Discrete Wavelet Transform Tilters for JPEG 2000 Standard", *WASET* , *International Journal of Signal Processing* Vo. 4, No.1 2008.
[4] K.Sayood, *Introduction to Data Compression*, 2nd edition, Academic Press, Morgan Kaufman Publishers, 2000.
[5] N. Ahmed, T. Natarajan and K. R. Rao, "Discrete cosine transforms*"IEEE Transactions Computer*" vol. C-23, pp. 90-93, Jan. 1974.
[6] Tsai, M. and H. Hung, "DCT and DWT based Image Watermarking Using Sub sampling," *in Proc. Of the 2005 IEEE Fourth Int. Conf. on Machine Learning and Cybernetics*, pp: 5308-5313, China, 2005.
[7] Rafael C. Gonzalez, Richard E. Woods *Digital Image Processing*, Addison Wesley publishing company – 2001.
[8] K. R. Rao, P. Yip, *Discrete cosine transform: Algorithms*, advantages, applications", Academic Press, San Diego, CA, 1990.
[9] S. Esakkirajan, T. Veerakumar, V. Senthil Murugan, and P. Navaneethan, "Image Compression Using Multiwavelet and Multi-stage Vector Quantization", *WASET International Journal of Signal Processing* Vol. 4, No.4, 2008.
[10] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. on Image Processing*, vol. 1, no. 2, pp. 205–220, Apr. 1992.
[11] C. Christopoulos, J. Askelof, and M. Larsson, "Efficient methods for encoding regions of interest in the upcoming JPEG 2000 still image coding standard," *IEEE Signal Processing Letters*, vol. 7, no. 9, Sept. 2000.
[12] I. E. G. Richardson, *Video Codec Design*, John Wiley & Sons, 2002.

[13] T. Acharya and P. S. Tsai. *JPEG2000 Standard for Image Compression: Concepts, Algorithms and VLSI Architecture*. New York: John Wiley & Sons, 2005.
[14] Mohammed M. Siddeq, *Image Restoration using Perception Neural Network*, M.Sc. – Thesis , Computer Science Dept. – University of Technology –Iraq. -2001.
[15] ADNAN KHASHMAN, KAMIL DIMILILER, "Image Compression using Neural Networks and Haar Wavelet", *WSEAS TRANSACTIONS on SIGNAL PROCESSING*, Vol. 4, No.5,2008.

Mohammed Mustafa Siddeq. He is birth date 1976 – Kirkuk – IRAQ. He is received his M.Sc. degree from University of Technology – Baghdad – IRAQ, and he is appointed as a lecturer and researcher at Software Engineering Dept.–Technical College –Kirkuk. He is published many researches about image compression in different journals inside and outside of IRAQ. He is published his researches at *Journal of Information and Computing Science – Academic Union – England*, and *Journal of Signal and Information Processing – Research Science Publisher – USA*. He has published works at *MathWork* Company in USA. He is professional in MATLAB language and Visual C++.NET He is area interesting: Digital Image Processing, Neural Network and Genetic Algorithm, Searching methods.

E-Commerce: True Indian Picture

Devendera Agarwal
Research Scholar, Shobhit University, School of CE&IT, Meerut, India
Email: dev_goel@in.com

R.P.Agarwal*, J.B.Singh* and S.P.Tripathi#
*Shobhit University, Meerut, India
Institute of Engineering Technology, Department of Computer Science, Lucknow, India
Email: prajanag@gmail.com, jbs.tomar@shobhituniversity.ac.in, spt@ietlucknow.edu

***Abstract*—This paper gives an insight of e-commerce and highlights the present scenario of e-commerce in India. It presents the surfing pattern of Indian public to give the critical review on truth of various reports being published from time to time. It also critically analyses the e-commerce with major focus on B2C e-commerce which involves e-tailing.**

***Index Terms*—e-Commerce, B2C, e-tailing, Indian Consumer, Trust**

I. Introduction

India is a country with rich historical heritage, the second most populous country and the most populous democracy in the world. It has achieved multifaceted socio-economic progress during the last 63 years of its independence and has once again emerged on world scenario as one of largest economies.

India [23] primarily being country whose economy encompasses the traditional village farming, finally accepted computer as an ally not foe. Computer growth in India moved in leap and bounds after much hyped Y2K problem. Since then India have make its presence felt worldwide in software industry with companies like Infosys, TCS, Wipro etc. grew exponentially. According to the International Monetary Fund, India's nominal GDP stood at US$1.3 trillion, which makes it the eleventh-largest economy in the world, corresponding to a per capita income of US$1,000. If purchasing power parity (PPP) is taken into account, India's economy is the fourth largest in the world at US$3.6 trillion. The country ranks 142nd in nominal GDP per capita and 127th in GDP per capita at PPP. With an average annual GDP growth rate of 5.8% for the past two decades, India is one of the fastest growing economies in the world.

According to a 2011 PwC report [24], in terms of PPP, India's GDP will overtake that of Japan in 2011 and by 2045, India's GDP will surpass that of the United States. Additionally, over the next four decades, India's average annual economic growth rate is expected to stand at about 8% and therefore, it has the potential to be the world's fastest growing major economy over the period to 2050. India has large numbers of well-educated people skilled in English language; India is a major exporter of software services and software workers.

The **Indian Information Technology industry** accounts for a 5.19% of the country's GDP and export earnings as of 2009, while providing employment to a significant number of its tertiary sector workforce. More than 2.3 million people are employed in the sector either directly or indirectly, making it one of the biggest job creators in India and a mainstay of the national economy.

Privatization of technical education also took place during this period and today India which churns out almost a million engineers every year. Public schools were not too far behind in imparting computer education as it was made compulsory from IIIrd standard itself. According to a study, titled "**India Urban consumer segment nationwide study 2009-10**", surveyed 19,178 respondents across 82 cities by Intel and IMRB, it was reported that computer penetration in urban India doubled in last three years from 19 per cent to 38 per cent and now nearly 28 million households have the PC in their houses.

The PC purchases have been driven by better education opportunity, internet connectivity and ease of working from home. Multipurpose usage of PC for gaming, watching videos and listening to music has also kicked off the sales of PC.

Technical education in India is governed by All India Council of Technical Education (AICTE) [4], which makes it compulsory to maintain ratio of 1 computer for every 4 seats in an engineering institute. It is also compulsory to have dedicated internet connectivity with bandwidth of 1 Mbps or more. Under normal circumstances any institute running for over 4 years must have at least 300 PCs. Moreover more and more institutes are offering students laptops at subsidized rates. School children have also started to demand computers for practice at home.

This gives an insight to why computer sales have surged in last few years. Computer and Internet are two essential components for e-commerce. Their increase has certainly had a positive impact on e-commerce growth in India.

II. E-Commerce Unleashed

The electronic commerce concept was developed in the 70's even though electronic commerce under the

doi:10.4304/jait.3.4.250-257

infant form of EDI or electronic data interchange has been existing since the late 60's with the invention of the first data networks [25].

Electronic commerce, commonly known as e-commerce or eCommerce, consists of the buying and selling of products or services over electronic systems such as the Internet and other computer networks. The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage. The use of commerce is conducted in this way, spurring and drawing on innovations in electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web at least at some point in the transaction's lifecycle, although it can encompass a wider range of technologies such as e-mail as well.

A large percentage of electronic commerce is conducted entirely electronically for virtual items such as access to premium content on a website, but most electronic commerce involves the transportation of physical items in some way. Online retailers are sometimes known as e-tailers and online retail is sometimes known as e-tail. Almost all big retailers have electronic commerce presence on the World Wide Web.

Electronic commerce is generally considered to be the sales aspect of e-business. It also consists of the exchange of data to facilitate the financing and payment aspects of the business transactions.

Originally, electronic commerce was identified as the facilitation of commercial transactions electronically, using technology such as Electronic Data Interchange (EDI) and Electronic Funds Transfer (EFT). These were both introduced in the late 1970s, allowing businesses to send commercial documents like purchase orders or invoices electronically. The growth and acceptance of credit cards, automated teller machines (ATM) and telephone banking in the 1980s were also forms of electronic commerce.

In 1990, Tim Berners-Lee invented the WorldWideWeb web browser and transformed an academic telecommunication network into a worldwide everyman everyday communication system called internet/www. Commercial enterprise on the Internet was strictly prohibited until 1991. Although the Internet became popular worldwide around 1994 when the first internet online shopping started, it took about five years to introduce security protocols and DSL allowing continual connection to the Internet. By the end of 2000, many European and American business companies offered their services through the World Wide Web. Since then people began to associate a word "ecommerce" with the ability of purchasing various goods through the Internet using secure protocols and electronic payment services.

With the advent of the World Wide Web (WWW), electronic commerce and especially company-to-consumer electronic commerce, is based on public networks such as Internet. Their main characteristic being that they are less expensive and widely accessible not only by corporations but also by the single individuals. There are many definitions of electronic commerce and much confusion there is about this term. For example Wigand [26] states that ***"Electronic commerce denotes the seamless application of information and communication technology from its point of origin to its endpoint along the entire value chain of business processes conducted electronically and designed to enable the accomplishment of a business goal. These processes may be partial or complete and may encompass business-to-business as well as business to consumer and consumer-to-business transactions".***

Zwass [27] defines electronic commerce as ***"The sharing of business information, maintaining business relationships, and conducting business transactions by means of telecommunications networks…Therefore as understood here, E-commerce includes the sell-buy relationships and transactions between companies, as well as the corporate processes that support the commerce within individual firms".***

A broader definition by Kalakota and Whinston [28] is: ***"E-commerce is associated with the buying and selling of information, products and services via computer networks today and in the future via any one of the myriad of networks that make up the Information Superhighway (I-way)".***

Internet also enables the marketers to easily reach the customers and promote their brands or products by offering vast product information and options. Electronic Commerce is the buying and selling of goods and services electronically by consumers or by companies via computerized transactions. E-Commerce has speeded up ordering, production, delivering, payment for goods and services by replacing manual and paper based business processes with electronic alternatives and by using information flow effectively in new and dynamic ways. At the same time, e-Commerce has reduced marketing, operational, production, and inventory costs in such a way that customer will also benefit indirectly.

Therefore, Internet is the technology [17, 18, 19] for e-Commerce as it offers easier ways to access companies and individuals at a very low cost in order to carry out day-to-day business transactions. Around the clock presence of companies on the Web gives competitive advantage to companies' businesses.

However, since the Internet is publicly accessible, data can be more easily intercepted, which seriously undermines the security of online transactions, as well as the privacy and confidentiality of the commercial exchange.

Moreover, the legitimacy and the trustworthiness of online vendors cannot be guaranteed as adequately as on a private network, because there is no control as to who will enter the system and how parties will authenticate themselves. Since users will often have the choice between a large numbers of different business partners and since the cost of switching from one vendor to another is negligible, it is imperative that online vendors stand out by addressing not only users' functional

business needs, but also their concerns in terms of security, confidentiality and trustworthiness.

For private users to adopt e-commerce, it is imperative that the benefits of using the new commercial medium (e.g. convenience, decreased transaction costs) significantly outweigh potential risks. Indeed, the private user's freedom to select appropriate vendors tends to be correlated with greater concerns regarding financial risk, privacy and trust. This can be accounted for by the fact that private users are more directly involved in the commercial exchange, since they are using their own equipment, giving sensitive information about themselves as individuals, and spending their own money.

A. *Types of E-Commerce*

The major different types of e-commerce are:

- **Business-To-Business (B2B)**

  B2B e-commerce is simply defined as e-commerce between companies. This is the type of e-commerce that deals with relationships between and among businesses. About 80% of e-commerce is of this type, and most experts predict that B2B ecommerce will continue to grow faster than the B2C segment.

- **Business-To-Consumer (B2C)**

  Business-to-consumer e-commerce, or commerce between companies and consumers, involves customers gathering information; purchasing physical goods (i.e., tangibles such as books or consumer products) or information goods (or goods of electronic material or digitized content, such as software, or e-books); and, for information goods, receiving products over an electronic network.

  It is the second largest and the earliest form of e-commerce. Its origins can be traced to online retailing (or e-tailing). Thus, the more common B2C business models are the online retailing companies such as Amazon.com, Drugstore.com, yahoo.com, rediff.com and indiatimes.com.

  B2C e-commerce reduces transactions costs (particularly search costs) by increasing consumer access to information and allowing consumers to find the most competitive price for a product or service. B2C e-commerce also reduces market entry barriers since the cost of putting up and maintaining a Web site is much cheaper than installing a "brick-and-mortar" structure for a firm. In the case of information goods, B2C e-commerce is even more attractive because it saves firms from factoring in the additional cost of a physical distribution network. Moreover, for countries with a growing and robust Internet population, like India delivering information goods becomes increasingly feasible.

- **Business-To-Government (B2G)**

  Business-to-government e-commerce or B2G is generally defined as commerce between companies and the public sector. It refers to the use of the Internet for public procurement, licensing procedures, and other government-related operations.

  A web-based purchasing policy increases the transparency of the procurement process (and reduces the risk of irregularities). To date, however, the size of the B2G ecommerce market as a component of total e-commerce is insignificant, as government e-procurement systems remain undeveloped.

- **Consumer-To-Consumer (C2C)**

  Consumer-to-consumer e-commerce or C2C is simply commerce between private individuals or consumers. This type of e-commerce is characterized by the growth of electronic marketplaces and online auctions, particularly in vertical industries where firms/businesses can bid for what they want from among multiple suppliers.

- **Mobile Commerce (m-commerce)**

  M-commerce (mobile commerce) is the buying and selling of goods and services through wireless technology i.e., handheld devices such as cellular telephones and personal digital assistants (PDAs).

  As content delivery over wireless devices becomes faster, more secure, and scalable, some believe that m-commerce will surpass wire line e-commerce as the method of choice for digital commerce transactions. This may well be true for the Asia-Pacific where there are more mobile phone users than there are Internet users.

This brief write-up with e-commerce discusses the evolution of e-commerce and how it has become an almost necessity in our day to day life. Frequent developments in technology particularly 3G and 4G in mobile will only add to the speed of growth of e-commerce.

Our major emphasis will be on B2C, as this type e-commerce is e-retailing or more common e-tailing. It involves the process of billing the end consumer. In our view this is where the true test of e-commerce takes place.

E-commerce requires monetary transaction, one single step where user hesitates to complete transaction. He already has heard so many electronic frauds [14, 15, 16], computers being hacked, passwords stolen etc., on the contrary truth is only (0.03%) of B2C transaction are fraud. 87% of fraud comes from online auctions (C2C e-commerce). If this myth can be broken it will prove to be a big leap in e-commerce.

In B2B and B2G e-commerce one party is a business house while other being a business house or a government organization. Both of them are well aware of threats of e-frauds which include manipulation of data records, hacking into organization systems, manipulation computer programs, unauthorized transfer of funds, failure of an e-transactions etc. Both business houses and government organizations have cyber security cells to maintain their computer system networks with the help of firewalls, proxy servers, antivirus software, white-list authorized wireless connections etc. They also have facility of legal advice to handle cyber crimes. On other hand a consumer in very small fish in sea who wants cheapest products usually falls in trap knowingly or unknowingly.

An e-commerce transaction requires a PC with internet connectivity and can be carried out from Home, Cybercafé or Office. At home we can assume PC to be safe, at office we have proper security of routers, ISA Server, Firewall and Antivirus so perhaps most secure place, while a cybercafé is perhaps the most susceptible place for a fraud.

It is essential to safeguard the interest of the consumer; it is he who will decide the fate of e-commerce in future. His trust has to be build; e-commerce will automatically grow with his trust and confidence. You can cheat him once, only to drive him away and not to trust e-commerce.

III. E-Commerce In India

Tall claims have been made about internet usage and e-commerce in India. Let's not go by the amount, as in B2B the numbers of transactions are negligible but amount involved is huge. B2B has always been here in form of EDI, so why there is so much fuss. It is B2C and C2C e-commerce which constitute majority of transactions of comparatively small amount.

The study, titled 'India Urban Consumer Segment Nationwide Study 2009-2010′ surveyed 19,178 respondents across 82 cities by Intel and IMRB, it was reported that Computer penetration in urban India has doubled in the last three years from 19 per cent to 38 per cent and now nearly 28 million households have the PC in their houses. The study also noted that youth in the age group of 18-25 are able to play a significant role as facilitators during the actual purchase of the PC.

The PC purchases [22] have been driven by better education opportunity, internet connectivity and ease of working from home. Multipurpose usage of PC for Gaming, watching videos and listening to music has also kicked off the sales of PC.

Study also found that, more first-time buyers are buying notebooks as their first computer. In 2006, only a mere 17 per cent of first-time buyers wanted to go for notebooks, while in 2009, the percentage of non-owners who wanted a notebook instead of a desktop PC doubled to 31 per cent.

One thing is sure that internet usage in India is increasing in leap and bounds. Many surveys [6, 7, 8, 9, 10, 11, 12, 13] show same is the case with e-commerce. Lets not go by the amount, as in B2B the numbers of transactions are negligible but amount involved is huge. It has always been there in form of EDI, so why there is so much fuss. It is B2C or C2C e-commerce where we have many transactions of small amount.

In this section first we find out the surfing pattern of Indians, to get the answer to two primary questions.

a) ***Are internet users really interested in e-commerce?***
if answer to above question is yes than to find
b) ***What we they buying on the internet and why?***

We found out the top 100 websites (hit ratio) [5] in India and classified them; the result is shown in Table-1.

Now let's explain each category and try to assimilate the necessary information on surfing behavior, and If possible, find out shopping pattern.

- **Portals:** A **web portal**, presents information from diverse sources in a unified way. Apart from the standard search engine feature, web portals offer other services such as e-mail, news, stock prices, information, databases and entertainment.

TABLE I.
SURFING PATTERN IN INDIA

S. No	Categories	Total
1	Portals	36
2	Advertisement	16
3	Entertainment	14
4	Social Networking	13
5	Search Engines	7
6	Internet Service Providers	6
7	Banks	4
8	E-Commerce	3
9	Encyclopedia	1
	Total	**100**

When further sub-categorized the picture came out as shown in figure-1. It is being observed that few portals are involved in e-commerce activities but keeping in mind Indian scenario, name of six websites is worth mentioning i.e. (Rediff India, Indiatimes, Ebay India & Sify) under ***General***, Shaadi under **Matrimonial** and Makemytrip under ***Travel*** Sub-categories. Remaining other offer various other services and hence are not worth mentioning.

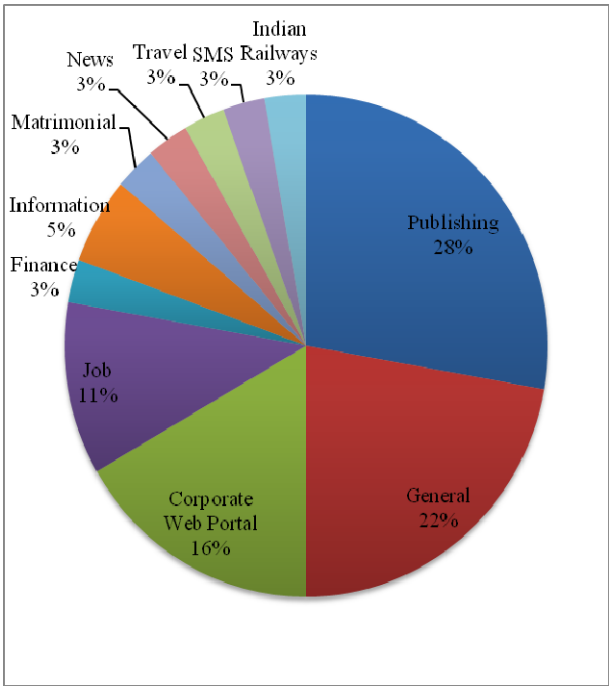


Figure 1. Sub-categories of Websites.

- **Advertisement:** Under these categories are those websites whom name one might have never heard of. They consist of those websites which generally pop on your screen when you visit other websites. They are basically advertisements on other websites and simply increasing their hit ratios Komli, Sulekha, Quikr India etc. are few of them.
- **Entertainment:** This category is yet another extension of portals, where whole emphasis is on entertainment only. It includes websites offering (Games, Music, Videos, Cricket scores etc.)
- **Social Networking:** It consist of latest in fashion sites meant for communication with friends, social causes etc. consist of common websites like Facebook, Twitter, Orkut, Bharatstudent etc.
- **Search Engines:** Perhaps one of most powerful tool on internet for actual working, no surprise, first two most popular websites being search engines Google and Google India, other are Bing, Ask, etc.
- **Internet Service Providers (ISP):** Again like advertisement category it comprises of websites which user actually doesn't visit himself but their hitcounter automatically hits when we open some other websites like websites of Hit counter (StarCounter, Conduit, etc.) and Domain Names at any spelling/typing mistake in name (GoDaddy, DomainTools, etc.).
- **Banks:** Banks are financial institutions an essential requirement for carrying out monetary transactions in e-commerce. There are actually three banks with one bank having two domain names (HDFC, ICICI and SBI). HDFC and ICICI are largest private banks in India; they are also pioneers of Internet Banking in India. Any net savvy user will certain have an account in these banks. Third State Bank of India (SBI) is largest and the oldest bank in India also offers net banking facility. Most business organizations have account in it and again finding its name in this category is no surprise.

  Presence of these names suggests that e-commerce is certainly present and making impact on India growth story.
- **E-Commerce:** Under these category we have placed those files that are strictly e-commerce (B2C) websites (irctc.co.in & amazon.com) and one being third party money transfer (paypal.com). Amazon is one of world most popular e-commerce websites, it is more likely that search engines direct user to Amazon rather user visiting this site for buying a product since payment is in Dollars & not in Rupees.

  PayPal is again facing problems with RBI guidelines so its presence in top 100 may be due to B2B transactions or some other reason and not from B2C transactions. Website of Indian Railway Catering & Tourism Corporation is a classical example of e-commerce growth in India and we discuss it in next section in detail.
- **Encyclopedia:** The single website of Wikipedia must attribute its presence to search engines, as it offers definition and history of each term being searched and it one of automatic choice for visiting the site.

In top 100 websites as per India's choice and conditions seven websites clearly can be termed as involved in B2C as they are selling products or services to Indian public in their own currency, and four websites providing payment gateway to carry out these transactions.

IV. TOP E-COMMERCE WEBSITES IN INDIA

Our next step was to individually analyze top e-commerce websites and find out what they have to offer and how much impact they make on India e-commerce growth. In the Table-II we have displayed these websites with their Indian Rank, World Rank and percentage of Indian audience visiting these websites.

TABLE II.
RANKING & %AGE OF INDIANS VISITING THESE WEBSITES

S.No	Website	Indian Rank	World Rank	%age Indian Audience
1	Rediff India	9	145	89
2	Indiatimes	12	173	78
3	IRCTC	36	574	98
4	eBay India	52	854	95
5	Sify	64	833	83
6	Shaadi	90	978	76
7	Makemytrip	98	1335	96

(Source: as per data available from alexa.com)

TABLE III.
CONTINGENCY TABLE FROM THE DATA OF TABLE II.

	Rediff	Indiatimes	IRCTC	eBay India	SIFY	Shaadi	Makemytrip	Total
Indian Rank	9	12	36	52	64	90	98	361
World Rank	145	173	574	854	833	978	1335	4892

Here Rediff India, Indiatimes and Sify can be grouped together in B2C category selling products of varied types Viz. mobiles, laptops, camera, watches etc., while eBay also deals in same but deals in C2C e-commerce.

We will be checking whether there exists a perfect combination between Indian Rank, World Rank and %age of Indian audience. To check this we use chi-square test [29, 30] of independence, considering the hypothesis as:

H_0: $PR_1 = PR_2$
H_1: $PR_1 \neq PR_2$

Where

PR_1 = Proportion of Indian Rank
PR_2 = Proportion of World Rank
Combined proportion of Indian Rank
= 361/4892
= 0.07

We compare the observed value of x2 with critical values of x2 and apply the rules of Hypothesis:

TABLE IV.
CONTINGENCY TABLE FROM THE DATA OF TABLE III.

f_0	f_e	$D=f_0-f_e$	D*D	$D*D/f_e$
9	51.57	-42.57	1812.20	35.1407
12	51.57	-39.57	1565.78	30.3623
36	51.57	-15.57	242.42	4.7009
52	51.57	0.43	0.18	0.0036
64	51.57	12.43	154.50	2.9960
90	51.57	38.43	1476.86	28.6381
98	51.57	46.43	2155.74	41.8023
145	698.85	-553.85	306749.82	438.9351
173	698.85	-525.85	276518.22	395.6761
574	698.85	-124.85	15587.52	22.3045
854	698.85	155.15	24071.52	34.4445
833	698.85	134.15	17996.22	25.7512
978	698.85	279.15	77924.72	111.5042
1335	698.85	636.15	404686.82	579.0754
			X_2=	1751.3349

$x_{2observed} < x_{2critical}$ => Accept the Null Hypothesis
and if
$x_{2observed} > x_{2critical}$ => Reject the Null Hypothesis

Now calculation the degree of freedom [31, 32] we get:

$\partial F = (r - 1)(c - 1)$
$\partial F = (2 - 1)(7 - 1)$
$\partial F = (1)(6)$
$\partial F = 6$
at $\alpha = 10\% = 0.10$
$x^2_{critical} / \alpha = 0.10$

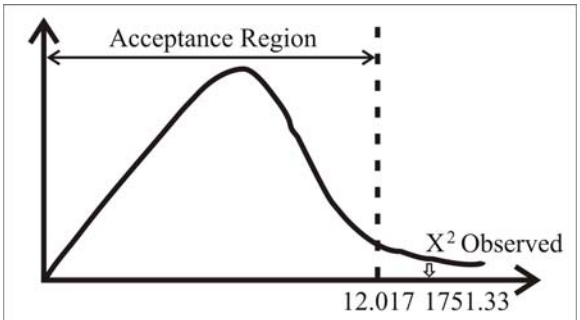


Figure 2. Sample Acceptance region

Since sample chi-square lies outside the acceptance region [33, 34] as shown in Figure-2 we reject the null hypothesis i.e.: the ranking of websites in India and the World does not have any correlation with each other.

TABLE V.
COMPARATIVE PRICES OF VARIOUS PRODUCTS OFFERED BY E-COMMERCE WEBSITES

S.No	Category	E-commerce Websites (Price in Rs.)					Lucknow Market
		Model No	Rediff India	eBay India	Indiatimes	Sify	
1	Mobile	Nokia	**7925**	8899	9711	9099	10670
2	Laptops	Compaq	32887	**30888**	33100	34999	33460
3	Camera	Canon	**6245**	8267	8095	6590	8320
4	Watches	FastTrack	**1895**	2240	2195	2200	2250

To find out, whether e-commerce in really useful other than convenience, time saving, etc., we visited all the B2C websites and found various similar products under different categories and also found their respective prices in surrounding area of living, making sure that we get the cheapest price and is also convenient to go and buy. To our surprise e-commerce also turned out to be the cheapest in all cases, with 7 days replacement guarantee and in most cases free postage and handling. The results of our finding are shown in Table V and Figure 3 represents the regression line showing relationship between Category & Prices.

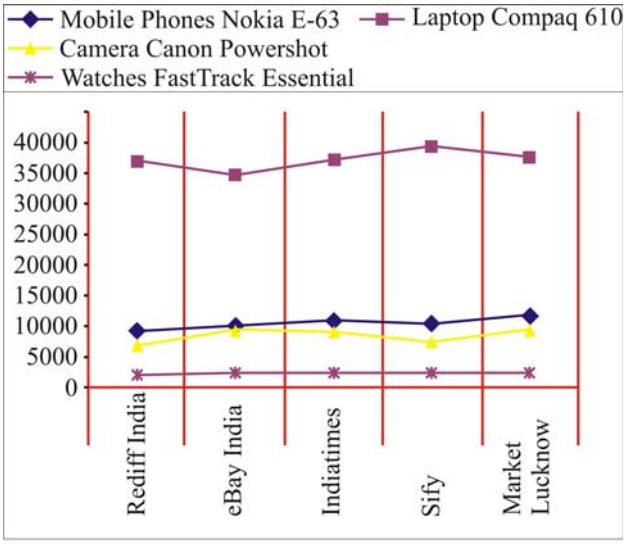


Figure 3. Regression Line showing relationship between Category &Price

Two other websites of Indian Railway Catering & Tourism Corporation (IRCTC) and Makemytrip both are related to travel industry i.e. booking of air and rail tickets and reservation in hotels. According to news nearly 40% of booked tickets are sold online. According to a report by IAMI (Internet and Mobile Association in India) 75% of total e-commerce business comes from travel industry and rest everything in B2C and C2C category contribute only 25%. Indian Railways fourth largest in the world carries 20 million passengers daily and IRCTC being official website for booking tickets is no surprise biggest contributor to e-commerce in India.

In India railways tickets can either be booked at railway station or online yourself/ through agents. Most stations offer booking service from 8 AM to 8 PM and tickets sold are cheaper then purchased online. But the shear amount of time it takes to get ticket booked tells upon the patience of any person. You require half/ full day leave to get tickets booked. Thus we see no reason for people to prefer to book tickets online.

V. Conclusions & Future Scope

We have seen that potential for growth of e-commerce in India is enormous. We have also seen that amount of interest that is there for travel industry is not seen in other services. Professional e-commerce websites [20, 21] are doing excellent job but what are the factors that are inhibiting users from purchasing online need to be ascertained.

The authors are working on the problem for last three years. They have tried to ascertain reasons in their papers [1, 2, 3] already communicated in referred journals.

In "Web Personalization of Indian e-Commerce Websites using Classification Methodologies" [2], authors have suggested how to identify fraud users by using Classification Methodologies using Bayesian Rules and generating cluster of users having fraudulent intentions.

In "TrFRA: A Trust Based Fuzzy Regression Analysis" [1], authors have focused on trust building factors in e-commerce websites. It focuses on fuzzy relationship between Trust and website related factors.

In "Trust Vs Complexity of E-commerce Sites" [3], authors have discussed how we need to tradeoff between Trust and Complexity so as not to drive away users from the website.

There are still various factors to be looked upon to cater to needs of the consumer who is the driving force for e-commerce his TRUST, CONVENIENCE, SECURITY are prime factors without which we will not be able to attain our goal.

Acknowledgment

The authors wish to thank management of Goel Group of Institutions, Lucknow for providing sponsorship in this endeavor.

References

[1] Agarwal, Devendera, et. al., "TrFRA: A Trust Based Fuzzy Regression Analysis", International Review on Computers and Software, Vol.5 N.6 November 2010, pp.668-670.
[2] Agarwal, Devendera, et. al., "Web Personalization of Indian e-Commerce Websites using Classification Methodologies", International Journal of Computer Science Issues, Volume: 7 Issue: 6, November 2010, pp. 18-21.
[3] Agarwal, Devendera, et. al., "Trust Vs Complexity of E-commerce Sites", to be published in International Journal of Scientific & Engineering, April 2012.
[4] All India Council of Technical Education (AICTE) website www.aicte-india.org
[5] ALEXA: The web Information Company website www.alexa.com
[6] Internet & Mobile Association of India (IAMAI) & Internet Market Research Bureau (IMRB), "I-CUBE 2008", Report by IMRB International, India, 2008
[7] Internet & Mobile Association of India, "I-CUBE 2009-10", Report by IMRB International, India, 2010.
[8] Internet & Mobile Association of India, "Cybecafe Users: E-commerce Activites", Report 2005.
[9] Internet & Mobile Association of India & Internet Market Research Bureau, "Consumer E-commerce market In India", Report 2006/07
[10] Internet & Mobile Association of India, "Annual Report 2006/07", Report 2007.
[11] Internet & Mobile Association of India, "Annual Report 2007/08", Report 2008.
[12] Internet & Mobile Association of India, "I-Cube 2007: Internet in India", Report 2007.
[13] Internet & Mobile Association of India & Internet Market Research Bureau, "I-Cube 2008: Internet in India", report 2008.
[14] Internet and Online Association of India (IOAI), "IOAI Survey: Ecommerce Security 2005", Survey Report 2005
[15] KPMG, "India Fraud Survey Report 2006", KPMG Forensic India.
[16] KPMG, "India Fraud Survey Report 2008", KPMG Forensic India.
[17] PEW/Internet, "Online Shopping", Survey Report 2008.
[18] Humpery Jonh, et.al., "The Reality of E-commerce with developing Countries", Report 2003

[19] Feldman Stuart, "The Changing face of E-commerce: Extending the Boundaries of the Possible", IEEE Internet Computing, 2000.
[20] Yong-Mi Kim, Suliman Hawamdeh, "The Utulization of Web 2.0 Functionalities on E-commerce Websites", Journal of Advances in Information Technology 2011, Vol2, No.4.
[21] K. Pragadeesh Kumar, Dr.N.Jaisankar, N.Mythili, "An Efficient Technique for Detection of Suspicious Malicious Web Site", Journal of Advances in Information Technology 2011, Vol2, No.4.
[22] Rastogi, Rajiv, "INDIA: Country Report on E-Commerce Initiatives", 2007.
[23] Wikipedia : The Free Encyclopedia (2011), "Introduction about India" Available: http://en.wikipedia.org/wiki/India
[24] Pwc Annual Report (2011), "Emerging Opportunities – Financial Services M&A in Asia 2011", Available at http://www.pwc.com/gx/en/mergers-acquisitions-industry-trends/survey/index.jhtml
[25] Ravi Kalakota, Andrew B. Whinston (1997), "Electronic Commerce: A Manager's Guide", Addison Wesley.
[26] Wigand, R. T (1997), "Electronic Commerce: Definition, Theory and Context, The Information Society", Vol 13, 1-16
[27] Zwass. V (1998), "Structure and Macro-Level Impacts of Electronic Commerce: From Technological Infrastructure to Electronic Market Places in Emerging Information Technologies" ed. Kenneth E. Kendall, Thousand Oaks, CA: Saga Publications.
[28] Ravi Kalakota , Andrew B. Whinston (1996), "Frontiers of electronic commerce", Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, 1996
[29] Greenwood, P.E., Nikulin, M.S. (1996), "A guide to chi-squared Testing", Wiley, New York.
[30] Corder, G.W., Foreman, D.I. (2009), "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach", Wiley.
[31] Eisenhauer, J.G. (2008), "Degrees of Freedom", Teaching Statistics, 30(3), 75–78
[32] Trevor Hastie, Robert Tibshirani, Jerome H. Friedman (2009), "The elements of statistical learning: data mining, inference, and prediction", Springer 2nd ed., 746 p.
[33] E. L. Lehmann (1997), "Testing Statistical Hypotheses: The Story of a Book". Statistical Science
[34] Lehmann, E.L.; Joseph P. Romano (2005), "Testing Statistical Hypotheses" (3E ed.). New York: Springer.

Devendera Agarwal is research scholar at Shobhit University, Meerut pursuing his Ph.D. in Computer Science. He has done B.Tech. in Computer Science from Mangalore University, Mangalore in 1993 and M.Tech. from U.P. Technical University, Lucknow in 2006.

Presently he is working as Director at Goel Institute of Technology & Management, Lucknow since 2008. His teaching areas are: Data Structures, Computer Graphics, Compiler Construction, Data Compression and Computer Programming. His areas of interest include Fuzzy Systems, E-commerce, Human computer interaction etc.

Dr. R. P. Agarwal is currently working as Vice Chancellor at Shobhit University, Meerut. He was former Vice Chancellor of Bundelkhand University, Jhansi. Prof. Agarwal has rich varied experience of teaching, research, development and administration. He has 41 years of teaching experience and has published more than 100 research papers in referred International and National Journals.

Dr. J.B. Singh is currently working Professor and Dean Students Welfare at Shobhit University, Meerut.
His teaching areas are Bio-Statistics, Statistics, Mathematical Modeling, Statistical Computing, Crop Yield Estimation, Operation Research and Neural Network.

Dr. S.P. Tripathi is currently working as Assistant Professor in Department of Computer Science at Institute of Engineering Technology, Lucknow. He is M.Tech. from IIT, Delhi in 1985 and Ph.D. from Lucknow University in 2005.
His teaching areas are Software Engineering, Database Management, Operating Systems, Data Mining and Computer Network.

Call for Papers and Special Issues

Aims and Scope

JAIT is intended to reflect new directions of research and report latest advances. It is a platform for rapid dissemination of high quality research / application / work-in-progress articles on IT solutions for managing challenges and problems within the highlighted scope. JAIT encourages a multidisciplinary approach towards solving problems by harnessing the power of IT in the following areas:

- **Healthcare and Biomedicine** - advances in healthcare and biomedicine e.g. for fighting impending dangerous diseases - using IT to model transmission patterns and effective management of patients' records; expert systems to help diagnosis, etc.
- **Environmental Management** - climate change management, environmental impacts of events such as rapid urbanization and mass migration, air and water pollution (e.g. flow patterns of water or airborne pollutants), deforestation (e.g. processing and management of satellite imagery), depletion of natural resources, exploration of resources (e.g. using geographic information system analysis).
- **Popularization of Ubiquitous Computing** - foraging for computing / communication resources on the move (e.g. vehicular technology), smart / 'aware' environments, security and privacy in these contexts; human-centric computing; possible legal and social implications.
- **Commercial, Industrial and Governmental Applications** - how to use knowledge discovery to help improve productivity, resource management, day-to-day operations, decision support, deployment of human expertise, etc. Best practices in e-commerce, e-commerce, e-government, IT in construction/large project management, IT in agriculture (to improve crop yields and supply chain management), IT in business administration and enterprise computing, etc. with potential for cross-fertilization.
- **Social and Demographic Changes** - provide IT solutions that can help policy makers plan and manage issues such as rapid urbanization, mass internal migration (from rural to urban environments), graying populations, etc.
- **IT in Education and Entertainment** - complete end-to-end IT solutions for students of different abilities to learn better; best practices in e-learning; personalized tutoring systems. IT solutions for storage, indexing, retrieval and distribution of multimedia data for the film and music industry; virtual / augmented reality for entertainment purposes; restoration and management of old film/music archives.
- **Law and Order** - using IT to coordinate different law enforcement agencies' efforts so as to give them an edge over criminals and terrorists; effective and secure sharing of intelligence across national and international agencies; using IT to combat corrupt practices and commercial crimes such as frauds, rogue/unauthorized trading activities and accounting irregularities; traffic flow management and crowd control.

The main focus of the journal is on technical aspects (e.g. data mining, parallel computing, artificial intelligence, image processing (e.g. satellite imagery), video sequence analysis (e.g. surveillance video), predictive models, etc.), although a small element of social implications/issues could be allowed to put the technical aspects into perspective. In particular, we encourage a multidisciplinary / convergent approach based on the following broadly based branches of computer science for the application areas highlighted above:

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the "Call for Papers" to be included on the Journal's Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. "Special Issue: Selected Best Papers of XYZ Conference".
- Sending us a formal "Letter of Intent" for the Special Issue.
- Creating a "Call for Papers" for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jait/.