Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 9 (19): 3531-3538, 2009
ISSN 1812-5654

Neural Network and Genetic Algorithm Based Hybrid Model for Content Based Mammogram Image Retrieval

[1]T.J. Jose and [2]P. Mythili
[1]Mangalam College of Engineering, Kottayam, Kerala, India
[2]Division of Electronics, Cochin University of Science and Technology, Cochin, Kerala, India

Abstract: In this study, an approach is described to content-based retrieval of medical images from a database provide a preliminary demonstration of our approach as applied to retrieval of digital mammograms. In the medical-imaging context, the ultimate aim of Content Based Image Retrieval (CBIR) is to provide radiologists with a diagnostic aid in the form of display of relevant past cases, along with proven pathology and other suitable information. We propose a new hybrid approach to content-based image retrieval. Contrary to the single feature vector approach which tries to retrieve similar images in one step, this method uses a two-step approach to retrieval. In the first step, we propose the use of a neural network called Self Organizing Map (SOM) for clustering the images with respect to their basic characteristics. In the second step, the GA based search will be made on a sub set of images which were having some basic characteristics of the input query image. We applied our approach to a database of high resolution mammogram images and show that this method radically improves the retrieval precision over the single feature vector approach. To determine whether our CBIR system is helpful to physicians, we conducted an evaluation trial with five radiologists. The results show that our system using genetic algorithms retrieval doubled the doctors' diagnostic accuracy. Moreover, this method is faster and has higher retrieval accuracy compared to the single stage methods.

Key words: Medical Image processing, medical image segmentation, region of interest, processing, CBIR, genetic algorithm, SOM, clustering, classification and tumor detection

INTRODUCTION

Breast cancer remains to be a leading cause of death among women in the developed countries. Currently mammography is the dominant method for detection of breast cancer. Mammography is an x-ray examination of the breast. It is used to detect and diagnose breast disease in women who either have breast problems such as a lump, pain, or nipple discharge, as well as for women who have no breast complaints. The procedure allows detection of breast cancers, benign tumors and cysts before they can be detected by palpation (touch). The sensitivity of mammography is approximately 90% (Mushlin *et al*., 1998). In spite of the technological advances in recent years, mammogram reading still remains a difficult clinical task. Some breast cancers may produce changes in mammograms that are subtle and difficult to recognize (Strickland and Hahn, 1996). Kopans (1992) have been reported that 10-30% of lesions are misinterpreted during routine screening of mammograms. Furthermore, it is very difficult to distinguish benign lesions from malignant ones in mammograms. As a result, between 2 and 10 women are biopsied for every cancer detected, causing needless fear and pain to women who are biopsied (Sickles, 1986). Due to the subtlety in the appearance of individual Micro Calcifications (MC), there is a significant risk that a radiologist may misclassify some cases in breast cancer diagnosis (Wei *et al*., 2005).

Training a neural network model essentially means selecting one model from the set of allowed models that minimizes the cost criterion. There are numerous algorithms available for training neural network models. Most of them can be viewed as a straightforward application of optimization theory and statistical estimation. Most of the algorithms used in training artificial neural networks are employing some form of gradient descent. This is done by simply taking the derivative of the cost function with respect to the network parameters and then changing those parameters in a gradient-related direction.

The Self-Organizing Map (SOM) is a subtype of artificial neural networks. It is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological

Corresponding Author: T. Justin Jose, Department of Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, Kottayam, Kerala-686631, India Tel: 0091 9946464172

properties of the input space. This makes SOM reasonable for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. The model was first described by the Finnish professor Teuvo Kohonen and is thus sometimes referred to as a Kohonen map. In this study, we use SOM neural network to cluster mammogram images into three distinct groups based on the characteristics of the background tissue of the mammograms. Kohonen (1995) described the structure of SOM network.

Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics inspired by evolutionary biology such as inheritance, mutation, selection and crossover (also called recombination). It is used in finding true or approximate solutions to optimization and search problems. GA is categorized as global search heuristics. Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary, as strings of 0 and 1 sec, but other encoding are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness) and modified (recombined and possibly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.

Images have always been used in medicine for teaching, diagnosis and management purposes. Now medical imaging systems produce more and more digitized images in all medical fields: visible, ultrasound, X-ray tomography, MRI, nuclear imaging, etc., Thus, for instance, Lund University hospital produces 15,000 new digital X-ray images per day. These images are very useful for diagnostic purposes. They are directly related to the patient pathology and medical history. However, the amount of images we can access nowadays is so huge that database systems require efficient indexing to enable fast access to images in databases. Despite the progress made in the general area of image retrieval in recent years (Bimbo, 1999), its success in biomedical thus far has been quite limited (Wong, 1998). Automatic image indexing using CBIR is one of the possible and promising solutions to effectively manage image databases (Smeulders *et al*., 2000). The CBIR from medical image databases does not aim to replace the physician by predicting the disease of a particular case but to assist him/her in diagnosis. The visual characteristics of a disease carry diagnostic information and oftentimes visually similar images correspond to the same disease category. By consulting the output of a CBIR system, the physician can gain more confidence in his/her decision or even consider other possibilities. The processing scheme adopted in the proposed system focuses on the solution of two problems. One is how to detect the ROI as suspicious regions with very weak background and another is how to extract features that characterize the suspicious regions. Many image collections contain few or no index terms. To search these collections, a set of techniques known as CBIR is used. The CBIR is a way to index or find a similarity between images in a database. The matching process between image search example and stored image content measures are complex and require sophisticated data management support. There are methods such as Fourier Transform, Hough Transform, Wavelet Transform, Gabor Transform, Hadamard Transform coefficients to be used as engine in CBIR system (Rui and Huang, 2000). Retrieval by image content has received great attention in the last decades. Several techniques have been proposed to the problem of finding or indexing images based on their contents (EI-Naga *et al*., 2002; Lamard *et al*., 2007; De Azevedo-Marques *et al*., 2008). Each method used has strong and weak points. All these traditional approaches to CBIR represent each image in the database by a vector of feature values (Flickner *et al*., 1995; Pentland *et al*., 1995).

A hierarchical learning framework for retrieval of relevant mammogram images has been reported in (EI-Naqa *et al*., 2004). A wavelet based Image retrieval method was proposed by (Lamard *et al*., 2007) which worked for all sort of images but to some extend it failed to retrieve the correct images. De Azevedo-Marques *et al*. (2008) a robust CBIR system with Relevance Feedback (RFb) for application in analysis of mammograms has been described which include several features related to the texture and distribution of breast density to index the images in the database, as well as techniques to incorporate the indication of relevance of the retrieved images provided by user. A classification approach assisted by CBIR (Yang *et al*., 2007) to improve the calcification accuracy in computer-aided diagnosis for breast cancer, reduced generalization error. A learning machine based framework for modeling human perceptual similarity for CBIR is reported by (EI-Naqa *et al*., 2004). But all these study have not mentioned about the speed of retrieval of masses. Persent earlier studies on functional

magnetic resonance image retrieval (Jose and Mythili, 2007) shows genetic algorithm will study fast for content-based image retrieval. In this study, we address both accuracy and speed of image retrieval. The results are more promising and accurate.

The proposed retrieval system is in principle very different and may helpfully complement existing diagnostic aids. We investigate the use of Genetic Algorithm Based CBIR system for digital mammograms. If we view the human observer as a classifier, then the aim of the CBIR system is to provide the observer with training-set examples that are close to his decision boundary. We expect such a facility to be useful in medical education and training as well. In this study we have proposed a method to provide the radiologist with a set of images from past cases that are relevant to the one being evaluated, along with the known pathology of these past cases. A library of relevant past images would assist radiologists to diagnose difficult cases in a better ways. The goal of the proposed CBIR is to obtain those mammograms that are similar in content to the query mammogram from a possibly very large mammogram database.

MATERIALS AND METHODS

This study was carried out in Digital Image Processing Laboratory of Mangalam College of Engineering during the period May 2007 to March 2009.

Overview of the proposed image retrieval framework: The proposed framework is illustrated with a functional diagram in Fig. 1. This framework will facilitate to search similar images in a large scale database with reasonable computational complexity.

Searching the entire database for similar images will increase the time required to retrieve similar images. Based on the character of background tissue mammogram images can be classified as F-Fatty, D-Dense-glandular and G-Fatty-glandular. Hence, the proposed model is divided in to two stages. In the first stage, the mammogram images present in the database is classified as F, D and G using SOM. In the second stage, the query image is obtained and the ROI is selected by the radiologist. As and when the ROI is selected the system acquire the length l and breadth b and then the class of the query image is identified using the same SOM network. Once, the class of query image is identified searches for similar suspicious region in the corresponding database with the same dimension of the ROI l×b will be followed.

Classification of mammograms: To classify the mammogram we have used the SOM Neural Network. SOM algorithm (Kohonen, 1995) is a neural network algorithm based on unsupervised learning. Basically it performs a vector quantization on the histogram of the images in the database and simultaneously organizes the quantized vectors on a regular low-dimensional grid. The block diagram of SOM classification is as shown in Fig. 2. Histogram of the image is chosen as input to the SOM since it is very simple to calculate. It was found that histogram feature was sufficient to classify images with 95% accuracy.

While finding, the histogram of a mammogram images, the continuous black background as well as the over exposed white regions will add considerable amount of error in the histogram output. To overcome this, the pixels which make the mammogram shape alone are considered for calculating the histogram. To achieve this,

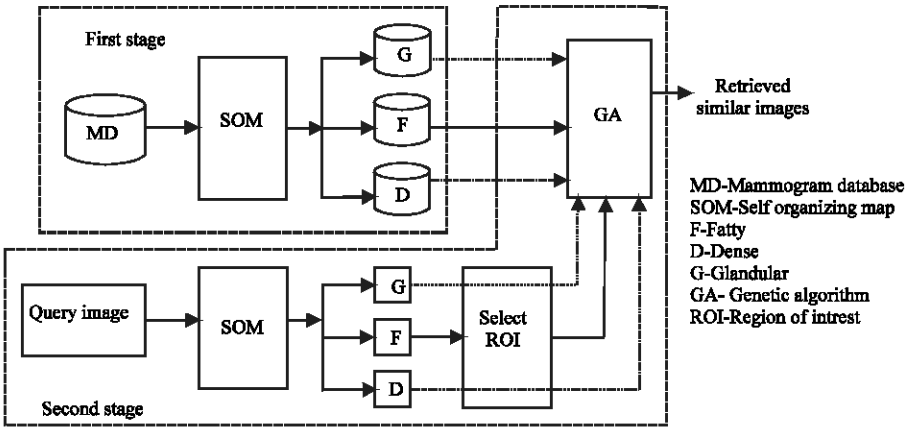

Fig. 1: Functional diagram of proposed model

as a preprocessing the pixels corresponding to the black background as well as the over exposed white regions were removed by using a suitable lower and upper threshold and the histogram corresponding to the remaining pixels were calculated.

Application of GA to this problem: In the second stage, after the classification of the query image and its ROI identified as shown in (Fig. 3a) every image from the specific class is taken and a random search is performed using GA over the entire image to locate the suspicious region matching with the ROI. The initial population for GA consists of chromosomes which represents x and y, the random position of ROI on the mammograms as shown in Fig. 3b. Binary chromosomes have been used for this purpose. A sample chromosome is shown in Fig. 3c. The corresponding decimal equivalent (x, y) coordinates are (119, 87). Roulette wheel selection and single point crossover (Goldberg, 1989) has been used. A fitness function proportional to correlation has been identified to evaluate the chromosomes as given in Eq. 1. Since, only the x and y are present in the chromosome the remaining coordinates can be calculated as (x+l, y),(x, y+b),(x+l, y+b) where l and b are length and breadth of the ROI which is fixed. Hence, ROI is always a rectangle with size l×b.

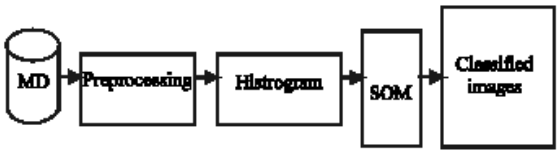

Fig. 2: Block Diagram of SOM Classification

$$\text{Fitness} = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}} \quad (1)$$

where, A_{mn} is a 2D matrix representing the ROI selected by the radiologist, $\bar{A}$ is the mean of A_{mn}. B_{mn} is a 2D matrix representing a portion of the image in the database bounded by (x, y),(x+l, y),(x, y+b),(x+l, y+b) and $\bar{B}$ is the mean of B_{mn}. The initial population consists of different x and y which correspond to different regions on a single mammogram in the database.

The outline of the CBIR algorithm: The complete outline of this model is as shown in Fig. 4. Initially the mammogram database is classified as and when a new mammogram is added. When the radiologist needs to find a set of images similar to the query image a ROI will be selected by him on the query image. Only one ROI can be selected at a time. If multiple ROI's are present then the search has to be repeated or the ROI's have to be merged in to a single ROI. After finding the class of the query

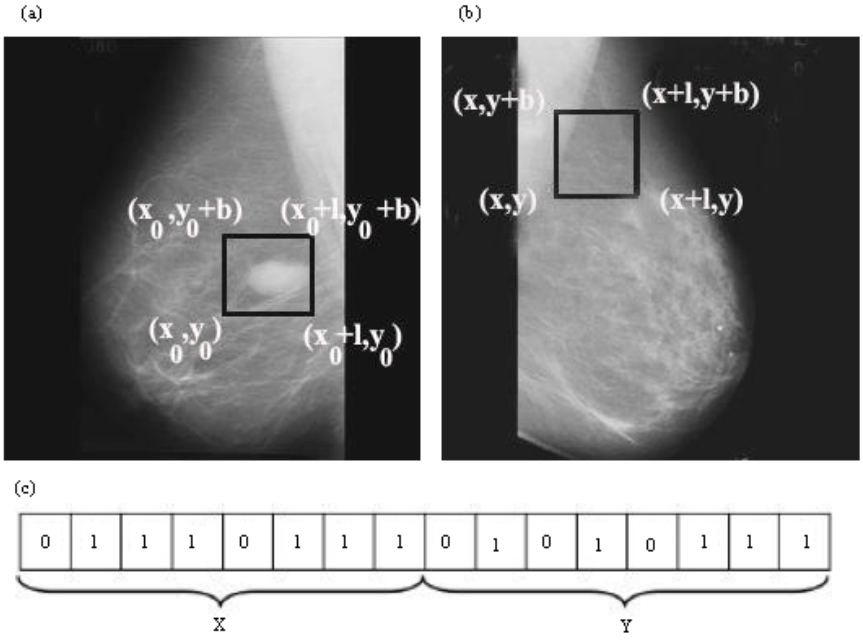

Fig. 3: (a) Position of ROI in query image, (b) Random position of ROI, (c) Structure of sample chromosome showing x, y coordinates

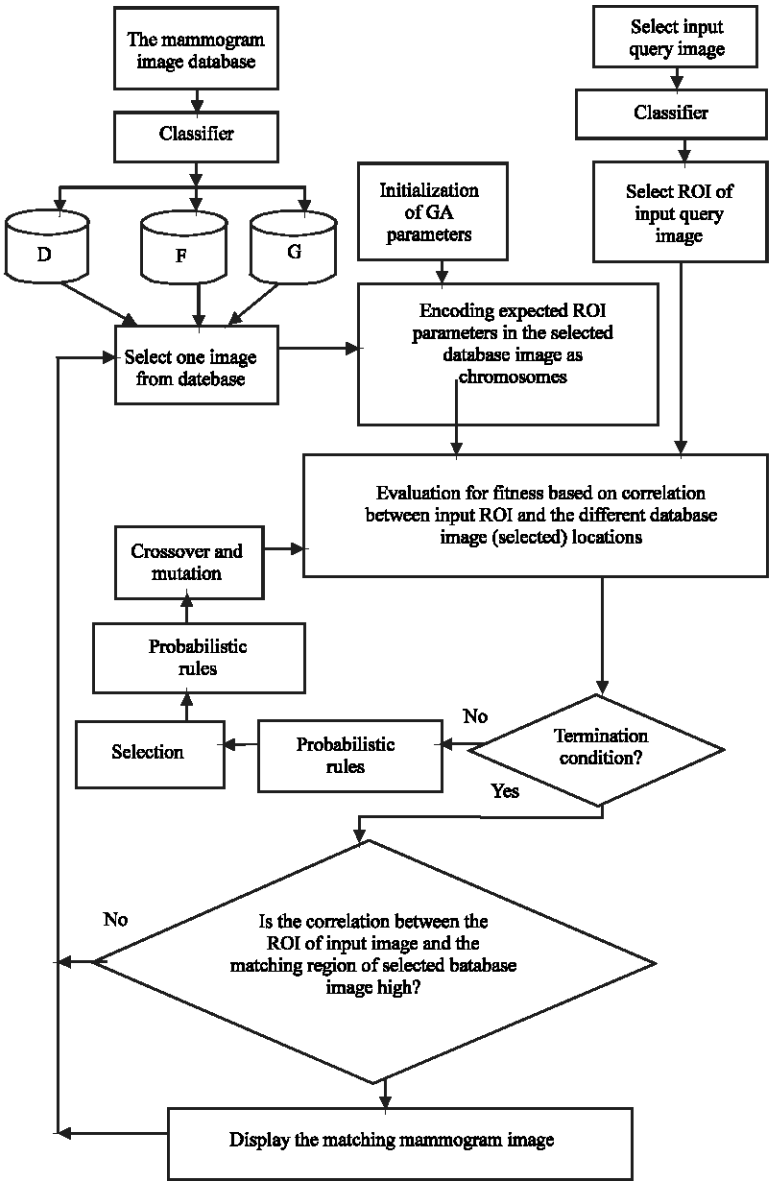

Fig. 4: Block diagram of the system

mammogram, each mammogram Di from that class database is selected and a random search is applied over it to find similarities. To do this random search a set of potential solutions are generated which represents the location of suspicious region. Each solution is checked for its closeness with the query ROI. This is done with the are performed to get a new set of solution. The above operations are repeated for g number of generations.

After g generations if the fitness value is grater than a threshold value then that image is considered as one of the solution image which is closer to the query image. This process is repeated to all the images in the specified database. After searching entire database the images are ranked based on their fitness and the best ones are given to the radiologist.

RESULTS AND DISCUSSION

The images for the proposed research was collected from the on-line digital database for screening mammography, located at the University of South Florida (These images were scanned from actual X-ray films taken of women being screened for breast cancer) and Mammographic Image Analysis Society's (MIAS) Mini-

Mammographic Database. While testing the system the images were down sampled to 256×256 and 512×512. Some of the images from the database were used as query images. A sample of known 30 F, 30 D, 40 G mammograms were taken for the analysis. This was classified using SOM. The classification accuracy was obtained as 97.6% based on few trails as shown in Table 1. Figure 5-7, show the mammograms automatically clustered by SOM texture property. If there are some misclassifications based on the Character of background tissue as well as according to the original manual classification denoted in the MIAS database, it can be negotiable in the sense that when such type of images become the query image the SOM will classify them accordingly.

Table 1: Classification accuracy of SOM

Trail	F	D	G	Percentage of accuracy
1	29	29	42	98.0
2	28	33	39	97.0
3	29	32	39	98.0

Mean= 97.6

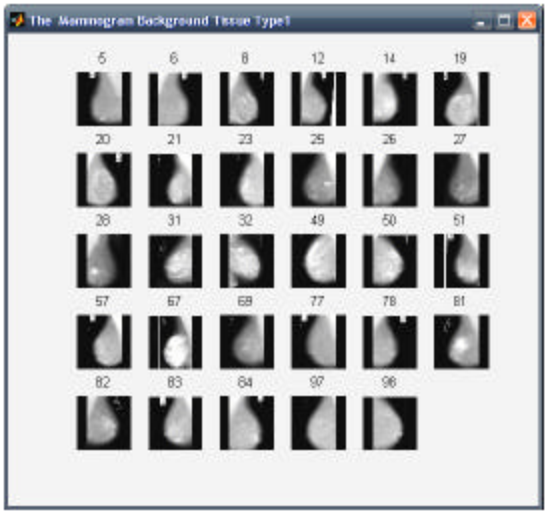

Fig. 5: Fatty images

Performance in terms of speed: The performance of the algorithms with the increasing number of mammogram images in the target database is shown in Table 2 and 3. The result has been compared with the conventional Sliding Window (SW) method. The result shows that our proposed method outperforms the SW method. The GA Parameters used for this study were population size = 25, No. of generations =10, mutation level = 0.2 and crossover rate = 0.7. The threshold level to identify the fittest elements was set to 0.8.

The proposed system has been successfully implemented and evaluated using MATLAB® on a normal 2G.Hz Pentium IV computer with 512 MB RAM. The proposed algorithm performed well even for large number of images. Table 3 shows that the proposed system

Table 2: Time taken to search with 256×256 images

	Time taken (sec)		
No. of images in the collection	GA	GA-SOM (ROI size = 32×32)	SW
30	25.34	10.35	50.7
60	67.98	27.46	100.0
100	101.32	41.81	170.2

Table 3: Time taken to search with 512×512 images

	Time taken (sec)		
No. of images in the collection	GA	GA-SOM (ROI size = 32×32)	SW
30	25.34	10.35	120.3
60	67.98	27.46	240.5
100	101.32	41.81	400.2

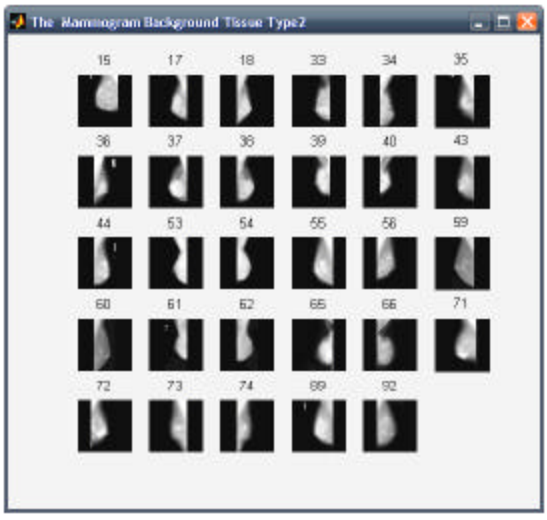

Fig. 6: Dense glandular

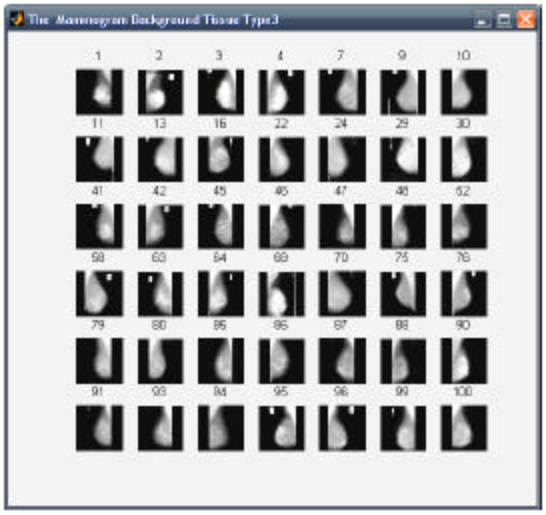

Fig. 7: Fatty glandular

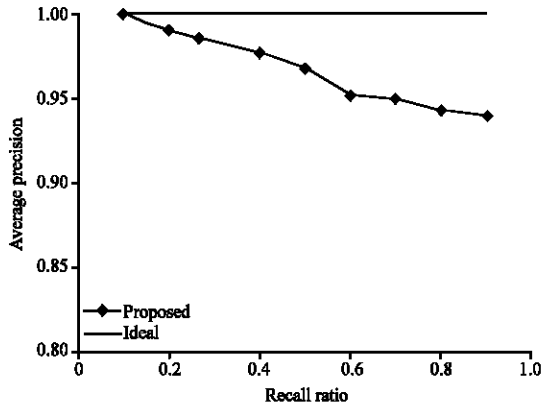

Fig. 8: Precision-recall study

will take constant time irrespective of the size of the image where as sliding window method will take more time depending upon the size of the image. The two stage network can provide 2~4 fold reduction in computation time as shown in Table 1 and 3.

Retrieval performance of proposed model: To evaluate the performance of the retrieval network, we used the so called precision-recall curves (Bimbo, 1999).The retrieval precision is defined as the proportion of the images among all the retrieved that are truly relevant to a given query image; the term recall is measured by the proportion of the images that are actually retrieved among all the relevant images to a query.

As a ground truth in calculation of the precision-recall curves, we considered an image to be truly relevant to a query if its corresponding observer Similarity Coefficient (SC) is larger than a pre selected threshold T. In present experiments, T = 7 was used. The observer SC study was carried out by a panel of five radiologist, who scored the similarity between each pair of ROIs based in their geometric distribution on a scale from 0 (most dissimilar) to 10 (most similar).The performance of the resulting overall network is summarized in (Fig. 8) using the precision-recall curves. From these result we can see that the proposed two stage model achieves a better performance than (Yang *et al*., 2007) and close to the ideal case (SW Method) which has very less error rate like an ideal system.

CONCLUSION

The proposed CBIR algorithm was developed and evaluated for retrieval of clinical mammograms containing suspicious region. The performance accuracy of the system has been tested with different kinds of input mammogram query images. The results demonstrated that this framework can be used effectively for retrieving visually similar mammograms from a database. It was demonstrated that a two-stage model can offer several advantages over a single stage one, including faster speed and retrieval accuracy. In our future work we will investigate the clinical benefit of using the developed model for computer-aided diagnosis.

ACKNOWLEDGMENTS

We thank Mr. Biju Varghese and Dr. Sajan Varghese, Director Mangalam Diagnostic and Research Centre, Kottayam for their whole hearted support by providing valuable suggestion and arranging radiologist for testing the algorithm. Also, a special thanks to Mr. Ragoth Singh, Lecturer, Mangalam College of Engineering.

REFERENCES

Bimbo, A.D., 1999. Visual Information Retrieval. 1st Edn., Kauffman Publishers, Morgan.

De Azevedo-Marques, P.M. and N.A. Rosa, 2008. Reducing the semantic gap in content-based image retrieval in mammography with relevance feedback and inclusion of expert knowledge. Int. J. CARS., 3: 123-130.

El-Naga, I., Y. Yang, P. Nikolas, Galatsana, N. Miles and Wernick, 2002. Content-Based image retrieval for digital mammography. ICIP IEEE, 3: 141-144.

El-Naqa, I., Y. Yang, N. Galatsanos, R. Nishikawa and M. Wernick, 2004. A similarity learning approach to content-based image retrieval: Application to digital mammography. IEEE Trans. Med. Imaging, 23: 1233-1244.

Flickner, M., H. Sawhney, W. Niblack, J. Ashley and Q. Huang *et al*., 1995. Query by image and video content: The qbic system. IEEE Comput., 28: 23-32.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. 1st Edn., Addison-Wesley, New York, USA., ISBN: 0201157675.

Jose, J. and T.P. Mythili, 2007. A model based tumor detection in brain MRI using genetic algorithm based image warping and template matching techniques. Proceedings of the International Conference on Information Systems and Technology, Dec. 14-15, MES College of Engineering, Kuttippuram, Kerala, India, pp: 120-125.

Kohonen, T., 1995. Self-Organizing Maps. 1st Edn., Springer, New York.

Kopans, D.B., 1992. The positive predictive value of mammography. Am. J. Roentgenol., 158: 521-526.

Lamard, M., G. Cazuguel, G. Enol e Quellec, L. Bekri, C. Roux and B.E. Cochener, 2007. Content based image retrieval based on wavelet transform coefficients distribution. Proceedings of the Cite Internationale France, Aug. 23-26, IEEE EMBS, pp: 4532-4535.

Mushlin, A.I., R.W. Kouides and D.E. Shapiro, 1998. Estimating the accuracy of screening mammography: A meta analysis. Am. J. Preventive Med., 14: 143-153.

Pentland, A., R.W. Picard and S. Sclaroff, 1995. Photobook: Tools for content- based manipulation of image databases. SPIE Storage Retrieval Image Video Databases II, 185: 34-47.

Rui, Y. and T. Huang, 2000. Optimizing learning in image retrieval. Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Jun. 13-15, Hilton Head Island, SC, USA., pp: 236-243.

Sickles, E.A., 1986. Mammographic features of 300 consecutive nonpalpable breast cancers. Am. J. Roentgenol., 146: 661-663.

Smeulders, A., M. Worring, S. Santini, A. Gupta and R. Jain, 2000. Contentbased image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Machine Intel., 22: 1349-1380.

Strickland, R.N. and H.L. Hahn, 1996. Wavelet transforms for detecting microcalcifications in mammograms. IEEE Trans. Med. Imag., 15: 218-229.

Wei, L., Y. Yang and R.M. Nishikawa, 2005. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans. Med. Imaging, 24: 371-380.

Wong, S.T.C., 1998. CBIR in medicine: Still a long way to go. Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries, Jun. 21, Santa Barbara, CA, pp: 125-131.

Yang, Y., L. Wei and M.R. Nishikawa, 2007. Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. IEEE Int. Conf. Image Process., 5: V1-V4.