An Efficient Content Based Image Retrieval for Normal and Abnormal Mammograms

[1]Vibhav Prakash Singh, [2]Ashim Gupta, [2]Shubham Singh, [1]Rajeev Srivastava
[1]Department of Computer Science & Engineering, IIT (BHU), Varanasi-221005, India
[2]Department of Electrical Engineering, IIT (BHU), Varanasi-221005, India
[1]vpsingh.rs.cse13@itbhu.ac.in, [2]ashim.gupta.eee12@iitbhu.ac.in, [2]shubham.singh.eee12@itbhu.ac.in, [1]rajeev.cse@iitbhu.ac.in

Abstract-**Diagnosis of breast cancer from mammograms is a crucial task. CBIR can support radiologists in their decision to retrieve similar mammograms out of a database to compare the past cases with current case. Pectoral muscle, labels, and artifacts, present in mammograms can bias the detection procedures. So extractions of these are an essential pre-processing step in the process of CAD. In this paper, an efficient content based image retrieval system is developed for normal and abnormal classes of mammograms. The pre-processing steps include are artifact suppression using CCL and Morphological operation, automatic pectoral muscle removal, and image enhancement using CLAHE. After pre-processing, we segment images using modified region growing algorithm, and using this segmented image, Histogram based statistical, Shape, Wavelet and Gabor features are extracted. Finally, images are retrieved using Euclidean distance similarity measure. Experiments on benchmark database confirm that the proposed segmentation and retrieval framework performs, encouraging than Fuzzy c mean, Ostu, and Region Growing based segmentation and retrieval approaches.**

Keywords-CAD; Pre-processing; Content Based Image Retrieval; Feature Extraction; Segmentation;

I. INTRODUCTION

Breast cancer remains the leading cause of death in women. According to World Cancer Research Fund, nearly 1.7 million new cases were diagnosed in 2012 [21]. In a developing country like India, breast cancer cases are expected to double by 2025. Mammography is currently the most widely used method for diagnosis and detection of breast cancer. Designing of computer processing algorithms can improve the detection and diagnostic information from mammograms. Although not perfect, these Computer Aided Diagnosis (CAD) systems can be used to provide second opinion to the radiologists. They could help the radiologists in the interpretation of the mammograms by finding visually similar mammograms out of a database to compare the past cases with current case.

Content Based Image Retrieval (CBIR) is a growing research area in computer vision, where we can retrieve visually similar images (as per query content) from a historical image database. These retrieved images may help radiologist to analyze the visual content with past cases images. The visual content (features) extracted from database images are used for indexing of images. At the query time, the same features are extracted and matched with entire database, and retrieve the closest images. CBIR system does not provide any diagnosis information of the retrieved images. It just retrieves similar images according to visual patterns, based on texture, color, shape or other important content.

For the effective segmentation of mammogram, it is necessary to detect the presence or absence of abnormalities in the mammograms. These abnormalities can be of various types such as micro-calcifications, Speculated, Circumscribed masses, or other miscellaneous types [2]. Since the image content of all these mammograms is not significantly different, it becomes difficult to segment the ROI and hence the results of the CBIR system are generally less accurate [3]. In this work we analyze different segmentation methods like fuzzy C-means and Otsu thresh-holding method and compare it with our modified Region-Growing method. Herein, different morphological operators were applied prior to some further enhancement using CLAHE. We also adjusted the gray-level histogram of the mammogram to reduce the False Positive and False Negative lesion segmentation of the mammograms. Rest of paper is managed as follows. Section II demonstrates the proposed methodology. Section III sheds some light on Results with some further Discussions. Section IV is concluding section.

II. METHODS AND MODELS

The proposed CBIR framework as shown in Fig.1, are divided into two parts, off-line feature extraction and on-line image retrieval. In the component of off-line feature extraction, images are firstly pre-process then feature of the images are extracted from database. Then these features are described as feature vectors of the images and store in a database. And at the time of on-line image retrieval, the user or the radiologist can submit a query image to the CBIR system to search for desired images having similar content. The system pre-process and represents this query with another feature vector with same process. The similarities between the feature vectors of the query and feature dataset (those of the images in the database) are then computed and stored. This similarity is computed using Euclidean Distance (ED) between the feature vectors. After this the system ranks the search results in increasing order of the Euclidean Distance and returns the images that are most visual similar to the query content.

978-1-4673-8507-7/15/$31.00 ©2015 IEEE

A. Pre-Processing

Pre-processing is used for improving the quality of the mammograms and make the feature extraction and segmentation processes more reliable. Many images in mini-MIAS database are affected by artifacts. Artifacts in the mammograms are of high intensity such as labels, opaque marker, and scanning artifact, which are necessary to remove in order to accurate and efficient segmentation. Moreover, as we know that visual appearance of pectoral muscle and dense tissue are same, leading to incorrect segmentations. So, we have to extract the pectoral muscle for avoiding the wrong segmentation.

1) Label and Artifact Suppression using CCL and Morphological Operation

We have used a very simple method of Connected Component Labelling (CCL) to remove the labels from the mammograms. In this method, the gray-level mammogram image is transformed to binary image with a relatively small Global threshold value of 0.09 (experimentally obtained). Then find the labels for the connected object using 8-neighbouring. Finally the region with the largest area (highest number of pixels) is extracted for further processing. After this, morphological operations are performed to remove all connected components (objects) that have fewer than max pixels. Other artifacts are also removed by morphological clean operation [5].

2) Pectoral Muscle Removal

The pectoral muscle is a triangular opacity across the upper posterior margin of the mammogram [6]. The texture of the pectoral muscle may also be similar to some abnormalities and may cause false positives for the retrieval of suspicious masses [7, 8].

Previously, pectoral muscle is suppressed using Hough transform through boundary approximation [9, 10]. The work by Ferrari *et al* [11] proposed a polynomial modelling of the pectoral muscle.

Here, we have used a new approach for segmentation and removal of pectoral muscle, in which first, we apply adaptive k-means algorithm for finding the regions of an image. Further find a seed point in the pectoral region and apply Region Growing algorithm with very small threshold. (This is done so that only the pectoral muscle is segmented and then consequently removed.)

3) Contrast Limited Adaptive Histogram Equalization

In mammogram image, masses appear brighter at the core and gradually darken as the image is traversed from the mass core toward the background. So, uniform distribution of contrast can resolve this problem up to some extent.

CLAHE tries to flatten the histogram to create uniform and better quality image. It is adaptive method derived from many histograms, and redistributes the lightness values of the image, and it is best for improving the local contrast [13]. Also CLAHE is easy in implementation and return high contrast image, hence used by the proposed work for image enhancement.

4) Image Segmentation using Modified Region-Growing

Image segmentation refers to extraction of region of interest (ROI) for further feature extraction. The various image segmentation techniques [14] can be categorized as region based methods, thresholding methods, clustering approaches: such as Fuzzy C-means Clustering, K-means clustering and texture based methods etc.

In this paper, we have implemented a region-growing based segmentation method. A region growing method seeks to add the neighboring pixels to the region around a pre-chosen seed if they satisfy certain intensity based criterion. In our algorithm, after removing pectoral muscle and applying CLAHE in pre-processing step, we choose the brightest pixel in the remaining mammogram as the initial seed point. In many cases it is found that this seed lies within the suspected lesion region. The similarity criterion used is based on intensity based gray level threshold. This means that the region is grown until all the pixels for which intensity difference with seed is less than the threshold are added. In order to prevent the region from growing indefinitely (as is the case with Glandular tissues), we have used another limiting criterion which reduces the threshold value. This is done using Haralick's GLCM based contrast value in which it was found that if contrast is high (>1.8) or low (<0.7) best results are achieved if the threshold value is set to 0.13. In cases where contrast lies between the aforementioned limits, the threshold value is set at 0.20. Fig.2 (a-d) shows the outcomes of pre-processing and segmentation steps. In which, Fig.2b reflects the removal of labels and pectoral muscles, Fig 2c show the enhanced image after CLAHE, and finally Fig.2d shows the result of proposed segmentation. Proposed approach easily detected the ROI part of mammogram.

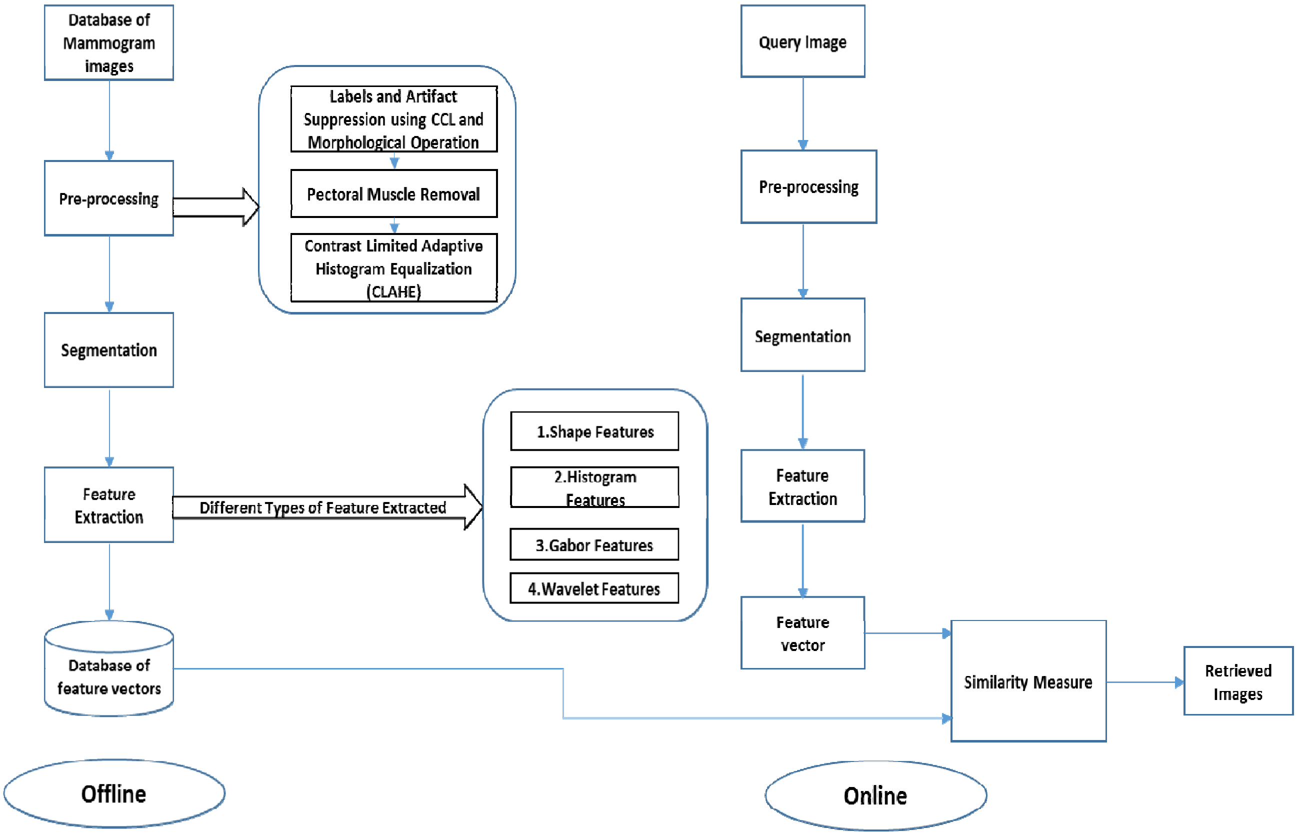
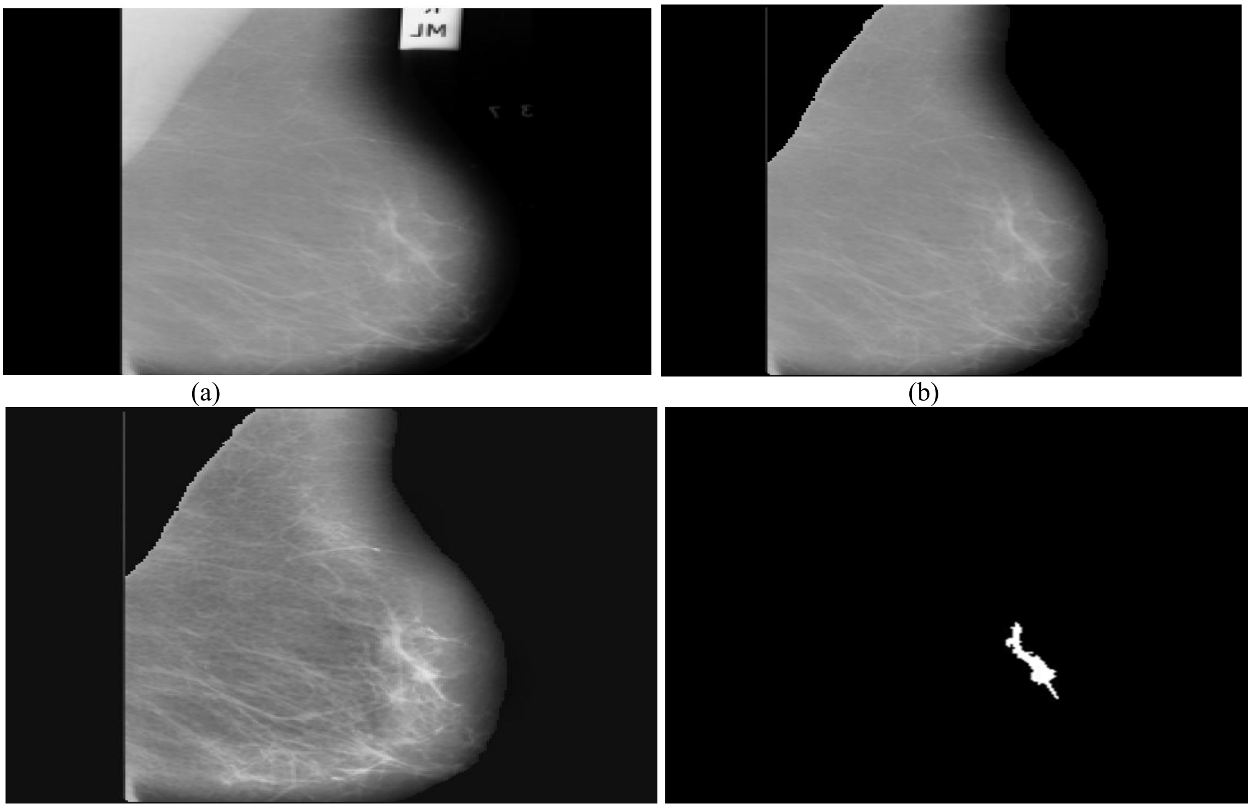


Fig.1 Working Model



(a) (b)

(c) (d)

Fig. 2a. Original Image, 2b. Image after artifacts, lables and Pectoral muscle removal, 2c. After enhancement, 2d. Segmented Image

An algorithm for modified region growing

i. Take pre-processed images, and find (x, y) = seed point, which is the point with maximum intensity value.

ii. Depending on the type of mammogram (fatty or dense), the threshold t for the termination of algorithm is determined. Calculate the Contrast of the mammogram using GLCM.

iii. If the Contrast is higher than 1.8 or less than 0.7, the threshold is taken as t = 0.13.
Else if the contrast lies between 0.7-1.8, the threshold value is set to t = 0.20

iv. Start region growing. The new point, from 8-neighboring pixels, is added to the segmented region if the criterion is satisfied.

v. Repeat the step 4 until the distance between seed point and the neighboring pixel is higher than threshold t.

Note that this technique is markedly different from those employed in most region growing algorithms wherein the difference between mean of the (already) segmented region and the new point is calculated. In this method, the difference is calculated between the (initial) seed point and the new point. This is done so as to prevent the segmentation of large area around the seed, and to segment only the suspected mass present in the mammogram.

B. Feature Extraction

Feature Extraction is the most important phase of image retrieval. Basically, features are the key attributes which are used for the discrimination of images. Here, this paper uses all categories of common features to obtain a hybrid feature set. The proposed hybrid feature set includes 8 geometric features (1-8), 6 histogram-based features (9–14), 18 Gabor features at 3rd level of decomposition (15-32) and two wavelet features (33-34) have been extracted.

1) Region and Shape Features:

Geometric or shape feature describe the geometric properties of the masses. In medical diagnosis for breast cancer detection, geometric features are essential to recognize any object, regardless of breast size. A simple geometric property includes image position, area, orientation, and centroid, etc. In this paper, for each image we have extracted 8 region properties that are Area, Euler Number, Centroid, Eccentricity, Perimeter, Filled Area, Convex Area, and Orientation.

2) Histogram Features

Statistical texture analysis is based on statistical properties of intensity histogram without considering spatial dependence. The histogram of the image gives brief statistical information about the image. Here, the useful features of the image that are obtained from the histogram, included; mean, which gives an average value of the intensity for an image. The variance tells the variation of intensity around the mean, skewness tells the symmetries of the histogram around the mean, the kurtosis is the flatness of the histogram. The entropy represents the uniformity of the histogram [1, 19].

Table I Features Details

Features	Dimension	Remarks
Shape Features	1-8	Shape features of segmented mammogram
Statistical Features	9-14	Based on first order statistics and other measures
Gabor Features	15-32	Mean square energy and Mean amplitude i.e., orientation θ extracted for the given images at three levels of Gabor wavelet decomposition in six orientations
Wavelet Features	33-34	Used for detection of transient changes in abnormalities. Mean and standard deviation are extracted of transformed image

3) Gabor Features

Gabor Filter is linear filter, widely adopted to extract texture features from the images for retrieval. Basically, Gabor filters are group of wavelets, capturing energy based features at a specific orientation and scale.

In order to extract textural micro patterns in breast mammograms, Gabor filters can be tuned with different angles and scales. The general function $g\ (x, y)$ of 2D Gabor filter can be represented as follows [18, 22]:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+jw(xcos\theta+ysin\theta\right] \quad (1)$$

where θ is orientation, σ_x and σ_y define the Gaussian envelope along the $X\ and$ Y-axes, and w is modulation frequency .

4) Wavelet Features

Wavelet gives an efficient representation of images, widely used for detection and retrieval [1, 18]. It has been used as a significant mathematical tool for decomposing a function in terms of its frequency and time components. The Discrete wavelet transform (DWT) captures both the frequency and spatial information of a signal. DWT analyses the image by decomposing it into approximation and detail coefficient via low-pass and high-pass filtering. Such decomposition is performed recursively on low-pass approximation coefficients at each level, till the desired outcomes. Here in this work, we have taken 3-level decomposition of low pass approximation of db2 wavelet. Afterwards Standard deviation and Mean of this transform coefficient are used as a feature vector. Table-I gives the complete descriptions of used features.

C. Mammogram Retrieval

All the above discussed features are extracted and store in database for indexing. For the searching and retrieval time radiologist put a query image, which features are extracted as same procedure. Afterwards, the similarity measures from every feature vector in the database to query image feature vector is calculated and stored. Finally, sort the similarity measure values in increasing order and searched chosen numbers of most relevant images.

Similarity measure, Euclidean distance (ED) is calculated as:

$$\mathrm{ED}(\mathrm{Q_Q}\,, \mathrm{F}_{DB}) = \sqrt{\sum(\mathrm{Q_Q} - \mathrm{F}_{DB})^2} \qquad (2)$$

where Q_Q is feature vector of query image and F_{DB} is feature vector of database images.

III. RESULTS ANALYSIS & DISCUSSIONS

1) Experimental Dataset

For our experimental analysis, we have used MIAS data base. The MIAS database has 322 images of mammogram; contain 115 images from abnormal and 207 images from normal classes. Here, Normal cases mean only healthy images, and Abnormal cases include both benign and malignant images [12].

2) Metrics Used For Performance Evaluations

The performance measures of the CBIR system quantitatively evaluated using precision and recall.

- *Precision*

Precision shows, how well system discriminate other images.

$$Precision\ = \frac{Number\ of\ Relevent\ Images\ retrieved}{Total\ number\ of\ retrievd\ images} \qquad (3)$$

- *Recall*

Recall shows, how well system recall the different images.

$$Recal = \frac{Number\ of\ Relevent\ Images\ retrieved}{Total\ number\ of\ relavent\ images\ in\ database} \qquad (4)$$

3) Image Retrieval Performance Analysis

The experiments are carried out, to randomly select 10 images (five from each category) as the query images with the number of retrieved images set as 40 for different comparative discussions. Here, we have taken only two classes normal and abnormal mammograms for retrieval. Firstly, we have calculated the average precision for random 10 query images, and number of retrieved images set as 10. This approach gives 67.5 % average precision for normal classes and 51% average precision for abnormal classes.

Fig-3 shows the snapshot of image retrieval for abnormal malignant query (image number-mdb014), where retrieval for this query is shown with corresponding image ids and it just gives the glimpse of effectiveness for the proposed work.

From the Fig.4, and Fig.5, it is clearly visible that average precision and recall of proposed framework are consistently encouraging for 40 images retrieval. So from comparative analysis with other segmentation approaches our framework confirms the effectiveness of the work.

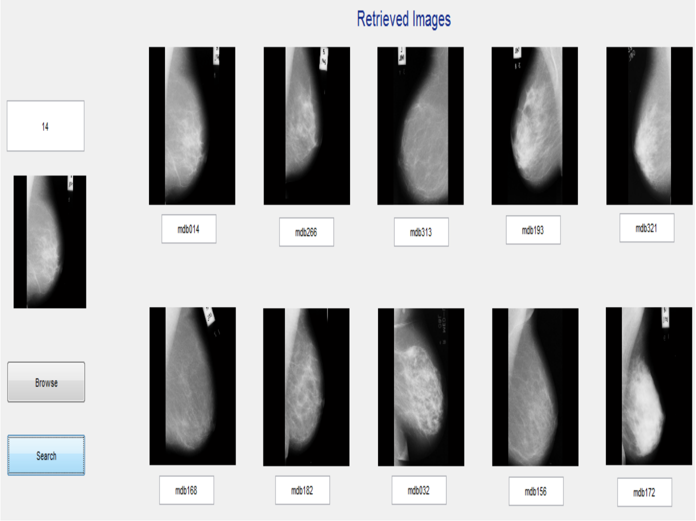


Fig.3. Image retrieval.

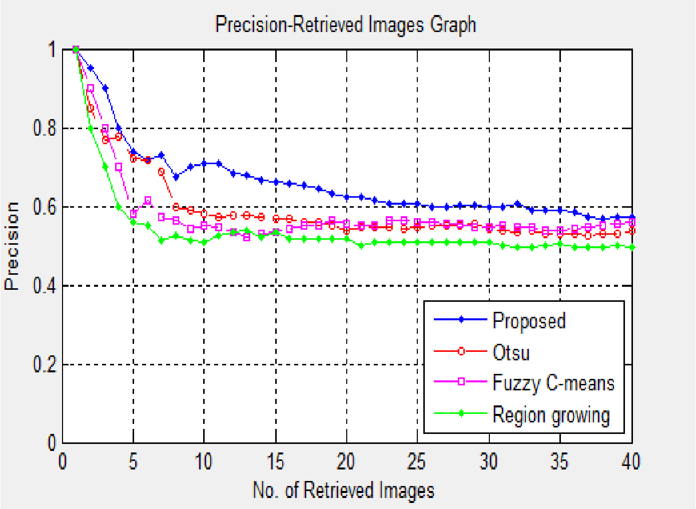


Fig.4 Average Precision with number of retrieved images

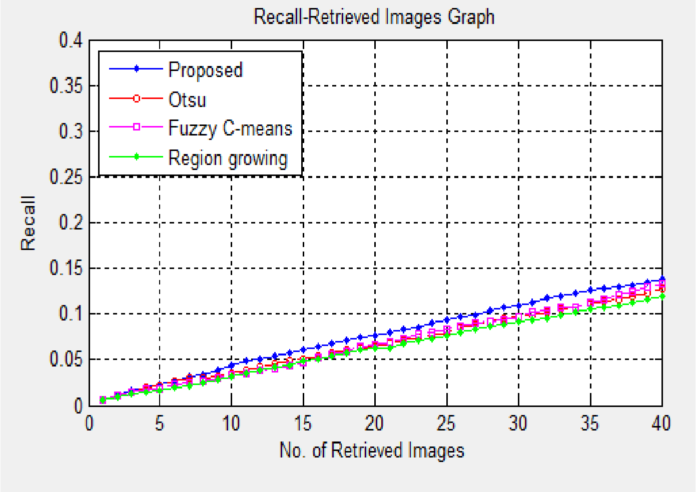


Fig.5 Average Recall with number of retrieved images

IV. CONCLUSIONS

As opposed to textual indexing, CBIR system, retrieve images based on visual content, which are highly benefited for early detection and diagnosis of cancer or non-cancer in mammograms. In this work, content based image retrieval (CBIR) system for two classes of mammograms from the MIAS database has been implemented. Here, first we automatically removed the labels, artifacts and pectoral muscles. Further, proposed modified segmentation approach confirmed the effectiveness of the work for image retrieval. This work, performances are significantly better than, Otsu, Fuzzy c means, and Region growing based segmentation approaches. In future, for enhancing the retrieval performance, we can use adaptive filters and dominating features (through feature selection methods).

REFERENCES

[1] Subodh srivastava et al." Quantitative Analysis of a General Framework of a CAD Tool for Breast Cancer Detection from Mammograms" Journal of Medical Imaging and Health Informatics Vol. 4, 1–21, 2014

[2] Sterns EE, "Relation between clinical and mammographic diagnosis of breast problems and the cancer/ biopsy rate," Can. J. Surg., vol. 39, pp. 128-132, 1996.

[3] R. Highnam and M. Brady, Mammographic Image Analysis, Kluwer Academic Publishers, 1999. ISBN: 0-7923- 5620-9.

[4] Jawad Nagi et al.,"Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms", IEEE EMBS Conference on Biomedical Engineering & Sciences, Kuala Lumpur, Malaysia, pp.87-92, 2010.

[5] Aswini Kumar Mohanty, Swasati Sahoo ,Arati Pradhan, Saroj Kumar Lenka, "Detection of Masses from Mammograms Using Mass shape Pattern", IJCTA, Vol. 2 (4), pp.1131-1139 , 2011.

[6] Eklund GW, Cardenosa G, and Parsons W: Assessing adequacy of mammographic image quality. Radiology 190 (2):297–307, 1994.

[7] Zhou C, Hadjiiski LM, Paramagul C, Sahiner B, Chan H-P, Wei J: Computerized pectoral muscle identification on MLO-view mammograms for CAD applications. Proc of the SPIE 5747:852–857, 2005

[8] R. Gupta, P.E. Udrill, "The Use of Texture Analysis to Delineate Suspicious Masses in Mammography", Phys. Med.Biol., VOL.40, PP.835-855 (1995).

[9] Karssemeijer, N., Brake "Combining single view features and asymmetry for detection of mass lesions". In: IWDM. (1998) 95–102.

[10] Kwok, S., Chandrasekhar, R., Attikiouzel, Y.: Automatic pectoral muscle segmentation on mammograms by straight line estimation and cliff detection. In: IIS Conference. (2001) 67–72 .

[11] Ferrari, R., Rangayyan, R.: Automatic identification of the pectoral muscle in mammograms. In: IEEE Transactions on Medical Imaging. Volume 23. (2004) 232–245.

[12] http://www.mammoimage.org/databases/

[13] Pisano, E.D., et al. "Image Processing Algorithms for Digital Mammography: A Pictorial Essay," Radio Graphics, vol. 20, no. 5, pp. 1479-1491, (2000).

[14] A. R. Domínguez and A. K. Nandi, Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. *Pattern Recognition* 42, 1138.

[15] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles", Vis. Res., Vol. 20, pp. 847–856, 1980.

[16] S. Zehan, B. George, M. Ronald, "Monocular Precrash Vehicle Detection: Features and Classifiers", IEEE Transactions on Image Processing, Vol. 15, Issue. 7, pp. 2019-2034, 2006.

[17] M. R. Turner, "Texture discrimination by Gabor functions," *Biology, Cybernetics, 55*, 1986, 71-82.

[18] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 7, No. 11, 1989,pp. 674–93.

[19] Chevrefits.C and Cheriet. F., Texture Analysis for Automatic Segmentation of Intervertebral Disks of Scoliotic Spines from MR Images, IEEE Transactions on information technology in biomedicine, vol.13 No. 4., 2009.

[20] Arfan. M., Jaffar and Latif .Classification and Segmentation of Brain Tumor Using Texture Analysis. Recent Advance in Artificial intelligence, Knowledge engineering and data bases.

[21] http://www.wcrf.org/

[22] Singh, Vibhav Prakash; Srivastava, Rajeev, Design & performance analysis of content based image retrieval system based on image classification using various feature sets, ABLAZE, Pages: 664 - 670, 2015.