ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Multi Query Image Retrieval System with Application to Mammogram Images

Simily Joseph*and Kannan Balakrishnan
Department of Computer Applications
Cochin University of Science and Technology
Kochi, India
simily.joseph@gmail.com, mullayilkannan@gmail.com

Abstract: Content Based Image Retrieval (CBIR) systems open new methods to efficiently manage large volumes of data. The proposed system accepts queries with single image and multiple images. The different queries are combined using logical AND. The features of the query image are compared with the features of the images stored in the database. The experimental result shows that the use of multiple queries has better retrieval performance than single image queries. This CBIR system is a general framework which can be used for different applications. In this study three kinds of texture features are used – Haralick texture features, Local Binary Patterns (LBP) and Haar wavelet features. The proposed system is tested with mammogram images.

Keywords: **CBIR, CBMIR,** image retrieval, feature extraction, breast cancer

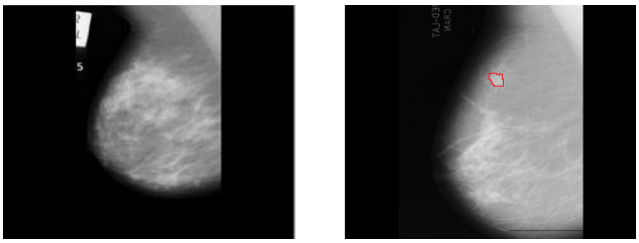
I. INTRODUCTION

With the popularization of digital information in the present era, the demand for effective storage and retrieval mechanism has reached an all time high level. The power of digital media and the advances in Information Technology offer enormous image archives and repositories for efficient information storage and processing. Content Based Image Retrieval (CBIR) systems retrieve desired images from vast collections of images based on low level features such as color, texture and shape. Traditional text based search engines exhibit a number of drawbacks. In those systems textual annotations are added manually, the process is subjective and incomplete. Text based search gives semantically similar images and content based search gives visually similar images [1]. The development of a Content Based Image Retrieval system could meet the storage and retrieval needs of information processing and management by solving the problems that exist in text based search engines [2]. CBIR systems retrieve images satisfying the conditions in the query, based on some low level features extracted from the image.

Nowadays Medical Image Retrieval is of great importance due to the huge volumes of clinical data. Radiologists can mark the pathology bearing region as Region of Interest and this region can be automatically compared with the reference images stored in the database. As similar cases share visually similar characteristics, the possibility of being normal or abnormal can be detected using the reference images. The integration of Content Based Image Retrieval approaches to daily clinical practice is helpful in clinical decision making. In medical field CBIR system allow the doctors to compare their diagnosis with the earlier findings in the respective domain in case of any doubt regarding the diagnosis. Thus by the use of human perception together with machine intelligence, best results can be arrived at, which is of great help for the physician. By inserting the given image into the retrieval system and there by obtaining the top most desired image the radiologist can arrive at conclusion in case of doubtful tissue in hand.

The proposed system can be used for retrieving mammogram images. Breast cancer is the second major cause of cancer death in women. Breast cancer is affecting the health and lives of millions and millions of women world over. Early detection is important for the complete cure of breast cancer. Fig.1 shows sample images of normal and malignant mammogram. Recently many inventions have been made by scientists to assist radiologists.

Masses and micro calcifications are not always cancerous even though they are considered as an early indication of breast cancer. Mammography is the best method for breast cancer detection with an average sensitivity of 80%. Microcalcification is an indication of a precancerous condition. Therefore the detection of clusters of microcalcification at an early stage is important for effective diagnosis and treatment. Microcalcifications are calcium deposits that are seen inside breast tissues. They appear as tiny white spots or as regions with size 0.1mm to 1mm [3, 4]. Mass detection is also an important factor for effective cancer treatment. It is more difficult to find masses than Microcalcifications; some reasons are the size, shape, change in density, low contrast, overlapping of non uniform tissues etc. In women with dense breast tissue the mammogram images are not visually clear, which make the diagnosis difficult. The wrong interpretation of mammogram may lead to unnecessary biopsy and further imaging. In such cases Content Based Medical Image Retrieval systems helps the radiologist in proper decision making.



Normal Mammogram Malignant Mammogram

Figure 1. Sample mammogram images

At present CBIR system developments mainly focus on single image queries. Apart from the conventional CBIR systems this work proposes a model which can handle

compound queries with multiple images. . Sometimes the mammogram will not be identical even if they belong to same diagnostic category. In such cases the information need can be expressed more accurately by multi query system. The advantage of using multiple image queries over single image query is that the former is more specific to the problem than the latter as the chance of expressing user's requirement is clearer. Moreover different stages of malignancy also can be retrieved by comparing the query images.

II. RELATED WORKS

A detailed survey of commercially available CBIR systems can be found in [5]. Recently several research papers have been published pertaining to mammogram image retrieval. Issam et al. [6] proposes a method in which the scores assigned by human observers are used to predict the perceptual similarity between mammogram lesions. This is based on a two stage hierarchical learning network. The extension of two stage hierarchical classifier with relevance feedback can be found in [7]. A method to identify the most relevant features that discriminate malignant and benign lesion is proposed by Joaquim et al [3]. A KNN based retrieval mechanism [4] uses structural characteristics and distributions of fibroglandular tissue and the size and shape of breast region. The performance of numerical classifier is improved with the use of similar retrieved images in [8]. Image retrieval for computer-aided diagnosis of breast cancer is proposed in [9]. Approaches towards the reduction of semantic gap can be found in [10]. The feedback from radiologist is used with query point movement techniques.

The textural features and the distribution of fibroglandular tissue are used for computing similarity. For query images of BIRADS breast density index 1, a small loss in the accuracy is observed. The increase in result with the increase in breast density indicates the significance of use of relevance feedback process. In multi view information based retrieval [11] the mediolateral oblique (MLO) and craniocaudal (CC) view of the query images are used. Eucledian distance and KNN method are used for comparison. This study reports better result than single view based retrieval.

All the above said works are based on single image queries. J. Tang et al. propose a method that uses multiple images for querying [12]. This method extracts one type of feature from one image and another type of feature from the other. The extracted features are combined and used for further similarity comparison. Single/multiple color extraction from multiple regions within an image is proposed in [13]. The multiple image examples are preprocessed and combined to form a single query in [14], each query images are evaluated independently and the results are combined. In [15] a linear combination of the distances of a test image to all images in the query image set was used. A two tier approach for multi image queries was proposed in [16]. The distinguishing feature of the proposed system is the use of the logical operation, AND for combining different queries. This multi query-multi feature system allows the users to express their information need using a combination of more than one query.

III. POPOSED IMAGE RETRIEVAL SYSTEM

The proposed system follows Query by Visual Example paradigm. Query by Visual Example retrieves images based on the low level features that are extracted from the given images. The retrieved images and the query images share some common characteristics. The presence of superimposed lesions in medical images sometimes leads to misinterpretation. Therefore images with known cases will be supportive for proper diagnosis. The architecture of the proposed CBIR system is shown in Fig 2.

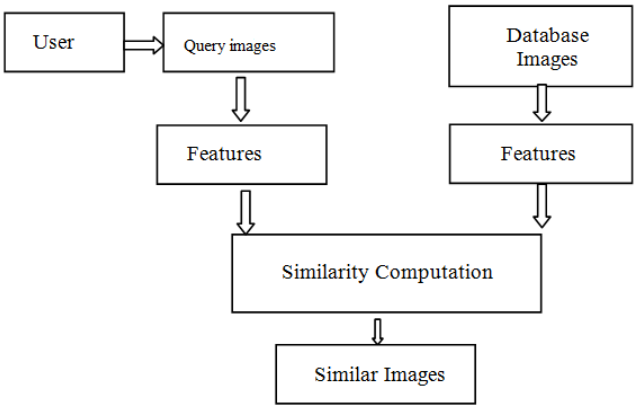


Figure 2. Architecture of the Proposed System

A. *Feature Descriptors:*

The images of different individuals show random variations due to noise present in the capturing devices and also due to the difference in physiology among individuals. Therefore the visual appearance of image does not contribute to image similarity or dissimilarity. To make each image distinct some measurable properties known as features are to be extracted. The image feature is represented as n–dimensional feature vector, $F = (f_1, f_2, \ldots\ldots f_n)$. The proposed system extracts texture features for finding the similarity between images. Texture is the visual patterns of homogeneity. It contains information about the structural arrangements of objects and their relationships. It is concerned with the spatial distribution of gray tones. As visual contents are not homogeneous, texture information are ideal for medical image analysis. The combination of three classes of features namely LBP, Haralick and Haar wavelet texture features reports good accuracy in mammogram image classification [17]. Therefore in this study we have used these three classes of features for finding image similarity.

Haralick features [18] are extracted from Greylevel Co-occurance Matrix $P_d[i, j]$, the matrix defined by first specifying a displacement vector $d = (d_x, d_y)$ and counting all pairs of pixels separated by d having grey levels i and j. GLCM contains information regarding the count of pixels having similar grey level values. The following five Haralick features are used in this study.

$$Contrast = \sum_i \sum_j (i-j)^2 P_d(i, j)$$

$$Energy = \sum_i \sum_j {P_d}^2(i, j)$$

$$Entropy = -\sum_i \sum_j P_d(i,j) \log P_d(i,j)$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{P_d(i,j)}{1+|i,j|}$$

$$Correlation = \frac{\sum_i \sum_j (i-\mu_x)(i-\mu_y)P_d(i,j)}{\sigma_x \sigma_y}$$

LBP (Local Binary Pattern) - a gray scale invariant when used for classification shows good performance [19]. LBP describes local primitives such as curved edges, points, spot, flat areas etc. To generate LBP code for a neighborhood, the weight assigned to each pixel is multiplied with a numerical threshold. The process is repeated for a set of circular samples. As a result the local binary patterns are said to be rotation invariant. Texture over a neighborhood of pixels can be defined as the joint distribution of the gray value of a central pixel of the neighborhood say g_c and gray value of circular pixels located at distance P.

$$T = t(g_c, g_0, g_1, ...g_{p-1})$$

The local texture pattern of a neighborhood can be obtained from the difference of central pixels and each pixel in the neighborhood. As the differences are independent this joint distribution can be factorized:

$$T \approx t(g_c)t(g_0 - g_c, ..., g_{p-1} - g_c)$$

To make this invariant against all transformations the signs of the difference are also considered and the overall luminance $t(g_c)$ is ignored as it does not contribute anything to texture analysis.

$$T \approx t(s(g_0 - g_c), ...s(g_{p-1} - g_c)$$

$$s(x) = \begin{cases} 1 x \geq 0 \\ 0 x < 0 \end{cases}$$

By assigning weight, this difference is converted to a Local Binary Pattern Code which is equivalent to the local texture. The following expression will generate 2^p LBP values for a neighborhood of pixels located at distance p.

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p$$

Where g_c is the gray value of a central pixel of the neighborhood. The process is repeated for a set of circular samples.

Wavelet coefficients are also used in this study. Wavelets are mathematical functions and are widely used in computer Vision applications. The simplest wavelet transform is Haar. Haar wavelet is discontinuous and similar to a bipolar step function. It provide an orthogonal basis for the space $L^2(R)$ .Consider the following function,

$$\Psi(t) = 1 \ for \ 0 \leq t \leq \frac{1}{2},$$

$$\Psi(t) = -1 \ for \frac{1}{2} \leq t \leq 1$$

$$\Psi(t) = 0 \ for \ all \ other \ cases$$

Haar wavelet over the interval [0, 1] can be defined as

$$\Psi_{k,n}(t) = 2^{\frac{k}{2}} \Psi(2^k t - n),$$

where $k \geq 0$ *and* $0 \leq 2^{k-1}$.

Haar scaling function can be defined as $\Phi(t) = 1 \ for \ 0 \leq t \leq 1, \Phi(t) = 0 \ otherwise$

These two functions together form an orthonormal basis for $L^2(0,1)$. In this study a 2 level decomposition is performed. A 2 level decomposition divides the grayscale image into 7 sub bands. The wavelet features from the low frequency components are extracted and used for similarity comparisons. The high dynamic range of features may wrongly affect the respective significance of features in image comparisons and classifications because features with larger value may have larger influences in decision making than features with small values. To overcome this problem features are normalized so that their value lies within (0, 1).

B. Multiple Image Query:

The similarity between the query image and the images in the database is obtained by calculating the distance of each feature of the query image and database image. In general, the distance function of query image and database image can be written as, $D(Q, P_i)$. Where Q is the query image, P_j is an image in the database. Q can be single image or multiple images. The distance will be minimum for highly similar images and will be maximum for highly dissimilar images. The distance measure used in this study is Euclidean distance. Application of logical operations simplifies the retrieval process. For AND operation, the resultant image should be similar to both query images. It implies that the distance of query image to the database images i and j should be minimum. Therefore by taking the minimum of maximum distance between Q and database image will meet the query condition. The distance function is MIN (MAX (D (Q_i, Pj))). Where D is the distance, Q_i, P_j represent the query images and database images respectively.

IV. EXPERIMENTAL RESULTS

The proposed system uses 312 mammogram images from MIAS database [20]. The images belong to three classes namely, 209 normal images, 42 malignant images and 61 benign images. Query images are randomly selected from the database. In the first stage of diagnostic process radiologist look for the presence of microcalcification clusters, masses, breast asymmetry based on size, density and shape, architectural distortion and nature of the background tissue. Therefore the CBIR system should work with all these kinds of abnormalities mentioned above. To satisfy this requirement, the proposed system extracts textural features from query images. When the query image is given its Eucledian distance with database images are calculated and sorted. The result of single image query with a malignant image is given in Fig: 3. Among the retrieved images, six are malignant and four are benign. The retrieved images resemble both the query images when the AND operation is used. The result is shown in Fig: 4. Among the images retrieved, eight are malignant and two are benign. This clearly shows the improvement in using multiple image queries.

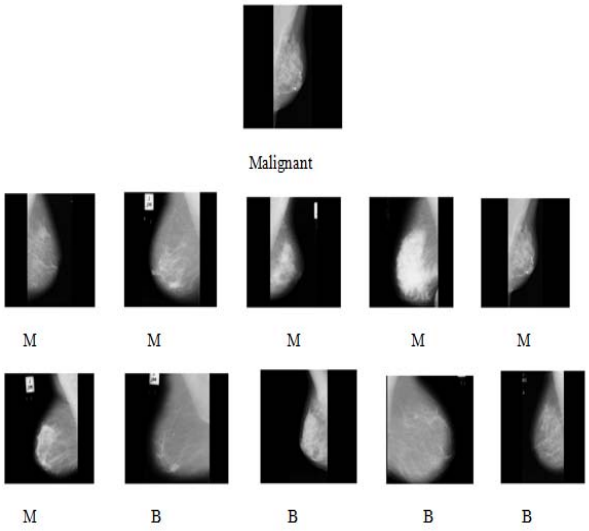


Figure 3. Result of Single image query-malignant image

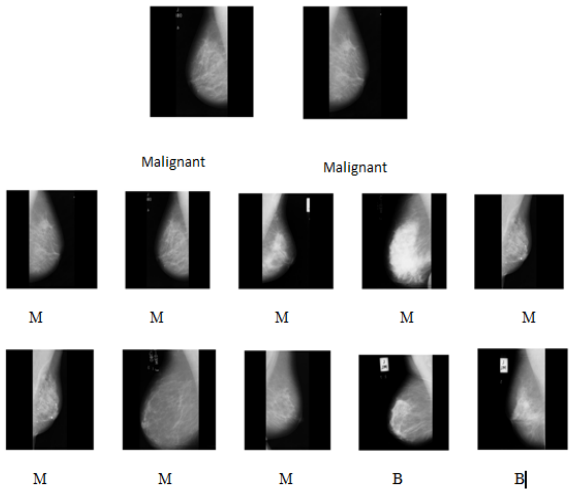


Figure 4. Result of AND operation between two Malignant Mammograms

The performance of the retrieval system is evaluated using precision and recall. Precision is the ratio of the number of relevant images retrieved to the total number of images retrieved, whereas recall is the ratio of the number of relevant images retrieved to the total number of relevant images in the database. The precision and recall values for multiple image queries are far better than single image query. The decrease in recall value for different query combination is due to the class imbalance problem. The number of normal images in database is high than malignant and benign images. Therefore the recall value also shows some difference

Table 1.Precision and recall for multiple image and single image queries

Query type	Accuracy	Malignant AND Malignant	Benign AND Benign	Normal AND Normal
Multiple image query	Precision	0.84	0.86	0.91
	Recall	0.40	0.27	0.086
Single image query	Precision	0.66	0.73	0.80
	Recall	0.238	0.180	0.057

V. CONCLUSION AND FUTURE WORK

In this paper, an efficient Content Based Medical Image retrieval System using multi query example is proposed .The system is tested with mammogram images. The features of both query images and database images are calculated and compared for similarity or dissimilarly. Depending on query condition a group of images are retrieved. The use of systems that support multi– image query promises better results than single image query system. The system assists radiologist in diagnosing cases where the difference between benign and malignant images are marginal. The retrieval accuracy can be increased by query refinement by collecting feedback from user which makes this more user oriented.

VI. REFERENCES

[1] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications – Clinical benefits and future directions". International Journal of Medical Informatics,73(1):1–23, 2004.

[2] W.C.Seng , S.H. Mirisaee, "Evaluation of a content-based retrieval system for blood cell images with automated methods". J. Med Syst 35 : 571-578, 2009

[3] J. C Felipe, M. X .Ribeiro, Elaine P M Sousa, Agma J M Traina, C. Jr Traina ," Effective Shape-based Retrieval and Classification of Mammograms" , ACM 1-59593-108, 2006

[4] Kinoshita, S.K., De Azevedo-Marques, P.M., Pereira, R.R., Rodrigues, J.A.H. & Rangayyan, R.M. Content-based retrieval of mammograms using visual features related to breast density patterns. Journal of digital imaging the official journal of the Society for Computer Applications in Radiology 20, 172-190 ,2007

[5] C.V. Remco, T.Mirela ," Content- based image retrieval systems: A Survey". Technical Report UU-CS—34,2000.

[6] El-Naqa I, Yongyi Yang, Galatsanos N.P, Nishikawa R.M, Wernick M.N ," Content based image retrieval for digital mammography". IEEE, 0-7803-7622-6/02, 2002

[7] El-Naqa I, Yongyi Yang, Galatsanos N.P, Nishikawa R.M, Wernick M.N,"A Similarity Learning Approach to Content-Based Image Retrieval: Application to Digital Mammography". IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 23, NO. 10, 2004.

[8] Yongyi Yang, Liyang Wei, Nishikawa R.M." Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis". Pattern Recognition 42 1126 – 1132, 2009.

[9] H.Jing, Y.Yang," Image retrieval for computer-aided diagnosis of breast cancer. Image Analysis & Interprtation": IEEE Southwest Symposium , pp.9 - 12 , 2010, doi: 10.1109/SSIAI.2010.5483930

[10] P.M. de Azevedo-Marques, N. A. Rosa, A.J. Machado Traina, C. Traina, S. K. Kinoshita R.M.Rangayyan, "Reducing the semantic gap in content-based image retrieval in mammography with relevance feedback and inclusion of expert knowledge". Int J CARS 3:123–130,2008

[11] R. Nakayama, H. Abe, J. Shiraishi,K. Doi," Evaluation of Objective Similarity Measures for Selecting Similar Images of Mammographic Lesions". Journal of Digital Imaging, Vol 24, No 1 pp 75-85, 2011

[12] J. Tang, S.Acton,An image retrieval algorithm using multiple query images, IEEE, 2003, pp. 193-196

[13] J. R. Smith, S.Fu Chang, Single color extraction and image query, International Conference on Image Processing (ICIP-1995), Washington, DC

[14] S.M.M. Tahaghoghi, A.T.James , H. E. Williams, Multiple Example Queries in Content-Based Image Retrieval, Lecture Notes in Computer Science, 2002, Volume 2476/2002, 227-241, DOI: 10.1007/3-540-45735-6_20

[15] Q. Iqbal,J. K. Aggarwal, Feature Integration, Multi-image Queries and Relevance Feedback in Image Retrieval, VISUAL 2003, Miami, Florida, Sep. 24-26, 2003, pp. 467-474

[16] H.C. Akakin, M. N. Gurcan, Content-based Microscopic Image Retrieval System for Multi-Image Queries, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, In press

[17] J.Simily , B.Kannan ," Local binary patterns, Haar wavelet features and Haralick texture features for mammogram image classification using artificial neural networks". Advances in Computing and Information Technology, Communications in Computer and Information Sciences 198: 107-114, 2011

[18] R.M.Haralick ,K.Shanmugham , D.Itshak , "Textural features for image classification". IEEE Transactions on Systems Man and Cybernetics SMC-3, No-6: 610-621, 1973

[19] D.Harwood , T.Ojala , A.M.Pietik, S.Kelman ,S. Davis ," Texture classification by center-symmetric auto-correlation, using Kullback discrimination of distributions". Center for Automation Research, University of Maryland CAR-TR: 678, 1993

[20] *J Suckling* et al ," The mammographic image analysis society digital mammogram database *exerpta medica". International Congress Series 1069, 375-378, 1994*

