Singaporean Journal Scientific Research (SJSR)
ISSN: 2231 - 0061Vol.2, No.2 pp.58 - 62

available at : : http://www.iaaet.org/sjsr

Content based mammogram retrieval usingGray Level Aura Matrix

Severin Wiesmüller[1], and D. Abraham Chandy[2]
[1] *Karunya University/Department of Electronics and Communication Engineering, Coimbatore, India*
Email: severin.wiesmueller@students.jku.at
[2] *Karunya University/Department of Electronics and Communication Engineering, Coimbatore, India*
Email: abrahamdchandy@gmail.com

Abstract—**Diagnosis of breast cancer in mammograms is also for specialists a difficult and error-prone task. A good opportunity to support radiologists in their decision is to find similar mammograms out of a database to compare the current case with past cases. In this work a complete content based image retrieval (CBIR) system for mass and calcification class mammograms has been implemented under usage of MATLAB®. The necessary feature extraction is realized on the basis of Gray Level Aura Matrix (GLAM). The normalized Gray Level Aura Matrices for the feature extraction is completely independent from shape and size of the region of interest (ROI) which gives a huge freedom for the user of the contentbased image retrieval system. Foundation for the built up database are 420 mammograms from the Digital Database for Screening Mammography (DDSM) database. The functionality of texture comparison using gray level aura matrix is demonstrated using the average precision of five retrieved images whereby the best result with 82.2% is reached with the nearest neighbourhood system and the quantization with eight allowed gray levels. Under these conditions the average precision of one retrieved image is 95%. Both results are better than comparable works.**

Index Terms—Content based image retrieval, mammogram, gray level aura matrix, image processing

1. INTRODUCTION

Accompanied with a growing amount of data the question of finding the most specific data for one special case becomes more and more important in nearly every application. Therefore developing new possibilities and algorithms for information retrievals is a fundamental field of research. Especially in conjunction with images a lot of benefits can be found outside of the usual text based image retrieval where a human has to define and write down the features of an image. One big problem going along with this activity is the huge expenditure of time which has to be spend from a specialist in the field, the image is from, to get proper metadata as keywords, tags or feature description. Additionally the problem of human subjectivity in producing this metadata makes a superior retrieval result very difficult. These handicaps can be avoided by using content based image retrieval. In this context “content” means data which is already in the image like texture, colour and so on. The challenge is here to find proper ways to extractdiscriminating content to find the most similar images for retrieval [1].

Many content based image retrieval systems, especially the one used in this project, can be described by the framework in Fig. 1 [2]. The CBIR process can be divided in an offline process and an online process. During the offline process all the selected images from the database pass through the feature extraction and all the features get saved. In this work the ROIs of every image in the database are identified because of its special coloured mark and for every ROI a feature vector based on the GLAM of the region is calculated and saved. In the online process a query image is given to the system [3]. It passes the same feature extraction and the features have to be compared using a similarity measurement with the saved features from the offline process. The result of this comparison is a list of all image files sorted by their level of similarity. In this work the similarity measurement is implemented as the measurement of the Euclidian distance between the feature vector of the image currently in the online process and all the saved feature vectors of the offline process, whereby the most similar images are characterized through the shortest Euclidian distance. Afterwards a chosen number of the best images get fetched from the database for visualization.One of the frequent diseases across the world is breast cancer and from this it follows that many people engage in breast cancer treatment and breast cancer detection. The most reliable method for detection of breast cancer is currently mammography [1]. Because of this huge amount of images the desire for an easy, fast and solid way of CBIR from mammogram databases is very large.

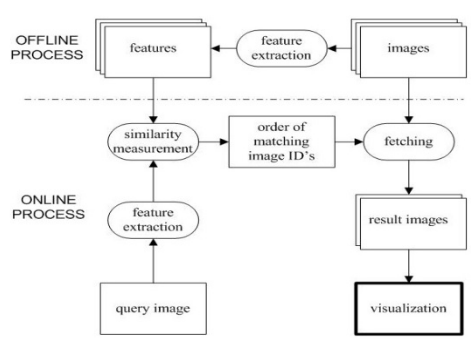


Figure 1.CBIR framework

This work’s focus is on content based image retrieval on the basis of similar texture extraction by GLAMs. The goals of this project are therefore the building up of a sample database out of DDSM images for testing the algorithm, the finding and cropping of ROIs according to colour marks in the images, the quantisation of gray levels in the ROI to minimize the size of GLAM without significant reduction of retrieval performance, finding out the best way of GLAM generation with different distances to find most fitting images, comparison of different pixel distance for GLAMs and testing and evaluation of the performance of the different possible settings.

2. RELATED WORKS

There are a lot of previous works in the field of CBIR for medical images. As the DDSM is a frequently used resource for researches in the field of mammography, in this paper only previous works with images from the DDSM are kept in mind. In Quellec et al.'s work [4] a CBIR method on the base of 250 to 500 images of the DDSM has been presented. With the method of wavelet optimization the reached efficiency, measured with the mean precision when five images are returned by the system was 70.91% for the classes normal, benign and cancer. Under similar conditions in Quellec et al.'s previous work [5] a precision of 82.1% was reached using the Dezert-Smarandache Theory. A different possibility of classify mammograms was used in de Oliveira et al.'s works [6], [7] where mammograms have been classified according to their density and lesion. The reached precisions in their works are between 80% and 90%. The most common breast abnormalities that may indicate breast cancer are masses and calcification and therefore many works are in the field of detection of these abnormalities as Bozek et al. present in their survey [8]. Purpose of this work is to retrieve mammograms of such abnormalities from DDSM. It was not possible to find works like this, so the above mentioned CBIR methods with other mamogram classes are used for precision comparison.

3. REALISATION

A. GLAM

An image can be modelled as a rectangular constitution S of m x n grids. Furthermore a neighbour-hood system $N=\{N_s,s\in S\}$ can be defined. At which the neighbourhood N_s is built from the basic neighbourhood E at site s. The basic neighbourhood is thereby a chosen structural element [9].

Aura Measure: [9] Given two subsets $A,B\subseteq S$, where $|A|$ is the total number of elements in A. The aura measure of A with respect to B is given in (1).

$$(\; , \; , \;) = \; | \; \cap \; | \quad (1)$$

∈

GLAM (Gray Level Aura Matrix): [9] Let N be the neighbourhood system over S and $\{S_i, 0\le i\le G\text{-}1\}$ be the gray level set of an image over S with G as the number of different gray levels, then the GLAM of the image is given in (2).

$$(\;) = \quad = \; (\; , \; , \;) \quad (2)$$

whereby $S_i=\{s\in S|x_s=i\}$ is the gray level set corres-ponding to the i^{th} level, and $m(S_i,S_j,\boldsymbol{N})$ is the aura measure of S_i with respect to S_j with the neighbourhood system $\boldsymbol{N}$.

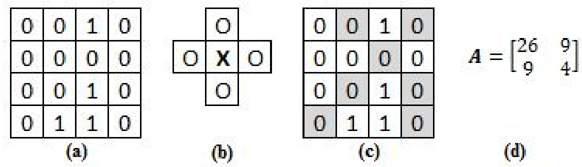


Figure 2 (a) A sample binary lattice S, where the subset A is the set of all 1's and B the set of all 0's. (b) The structural element of the neighbourhood system. (c) The shaded sites are the sites who are involved for building $m(S_1,S_0,N)$. (d) The corresponding GLAM

The aura of A with respect to B characterizes how the subset B is represented in the neighbourhood of A. The GLAM of an image measures the amount of each gray level in the neighbourhood of each gray level. As an example, the GLAM for the image shown in Figure 2 (a) is shown in Figure 2 (d), which is calculated using the structural element of the four-nearest-neighbourneighbourhood system as shown in Figure 2 (b).

Because of the 16 bit resolution of the original image the GLAM would be a matrix with a maximum size of 65536 x 65536. To reduce the size of the matrix and the necessary time for the retrieval the ROI has to be quantized before the GLAM generation. As discussed in the results the smallest possible number of allowed gray levels without loss of performance is eight. The result is a matrix with 64 entries which is transformed to a feature vector with 64 entries and normalized for the feature comparison. Because of the normalization the GLAM gets independent from the size of the ROI.

B. Mammogram Dataset

The mammograms are derived from the Digital Database for Screening Mammography (DDSM). The aim of the DDSM is to provide researchers a large set of mammograms to evaluate and compare the performance of computer-aided detection algorithms. It contains 2620 cases with four view mammography screening exams and corresponding data [10]. So every case includes the digitized copies of four mammograms (craniocaudal (CC) and mediolateral-oblique (MLO) view of each breast) in the true lossless JPEG format. The corresponding data contains the date of study, the patient age, the American College of Radiology breast density, the date of digitization, and the size and scanning resolution for each image. A separate OVERLAY file is provided for each image that contains marked lesions. These OVERLAY files contain the assessment, subtlety and pathology, and a description and chain code for each lesion [11]. As the handling with images in the lossless JPEG (LJPEG) format is not possible with MATLAB® the used mammograms are derived from the IRMA project in the user-friendly PNG format with 16 bit resolution. These images are the processed images from the DDSM courtesy of TM Deserno, Dept. Of Medical Informatics, RWTH Aachen, Germany. For this work the images have been sorted according to their corresponding information into three types of cancer classes (circumscribed, microcalcification and spiculated) and one normal class without any abnormality. Example mammograms for the four classes can be seen in Figure 3.

Images from the circumscribed class are characterized by some circumscribed mass which is evident inside of the mark. In the microcalcification class small calcifications are evident and marked whereby in the spiculated class a lump of tissue with spikes or points on the surface is evident and in pictures from the normal class no abnormality is marked by an oncologist.As it can be seen in the circumscribed example, it is also possible that more than one abnormality is marked in one image. In this case different colours for marking are used for the several regions of interest.

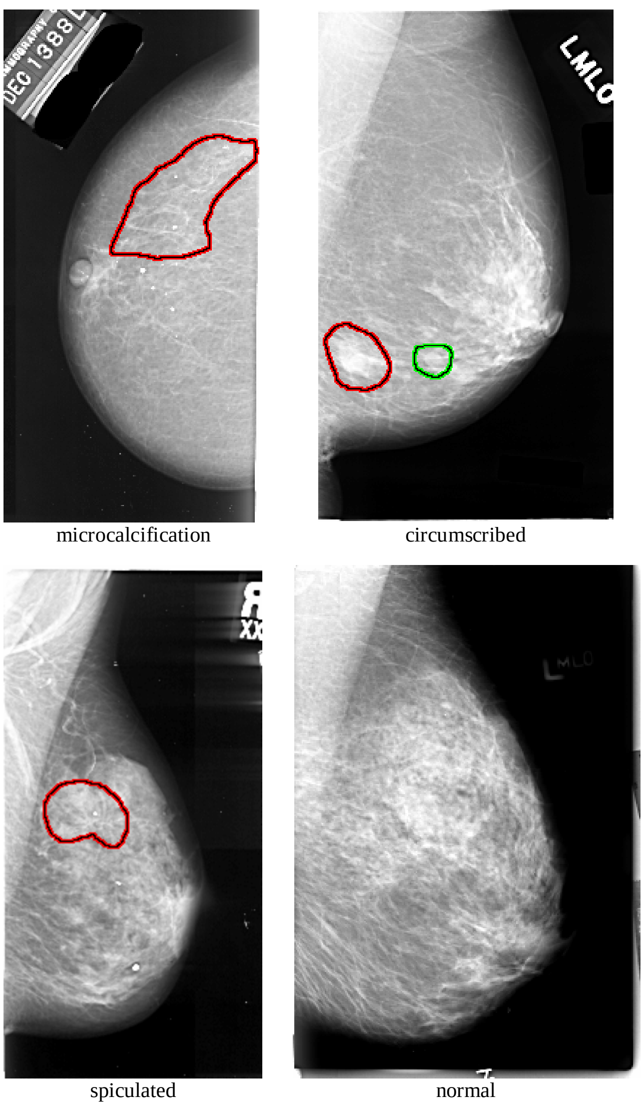


Figure 3 Examples for mammograms from the DDSM

In fact for this work two types of images are used. For the actual GLAM generation unmarked originals in high resolution were derived from the IRMA project and for finding the region of interest marked thumbnails were derived from the DDSM. To mention is, that in the images from the normal class some ROIs were used which looked suspicious for a regular untrained eye.

C. Region of Interest Cropping

In the thumbnail images the colours red (RGB: 255/0/0), green (0/255/0), blue (0/0/255), cyan (0/255/255), purple (164/4/255) and pink (255/0/255) are possible as marking colours. Through an extraction of the RGB colour information a mask for each of these colours can be build and the position of the mark in the image can be easily identified. Based on this information the position and size of the cropping rectangle is calculated through the extreme values of the mask in x and y direction and the scale between thumbnail and original image. Since every shape of image can be taken as basis for GLAM generation this calculated rectangle is cropped out of the original image and saved as one region of interest.

D. Mammogram Retrieval

For the image retrieval a query image runs through the process of ROI cropping, quantization and feature vector extraction. Afterwards the Euclidean distance from every feature vector in the database to this feature vector is calculated and saved. Finally the shortest distances are searched and a chosen number of most relevant images are retrieved. For testing the CBIR algorithm in this work 25 query images of each class are passing this process and the five most similar images are retrieved. Then every retrieved image is rated as correct or incorrect depending on its class affiliation to form a statistical base for result comparison.

4. RESULT AND DISCUSSION

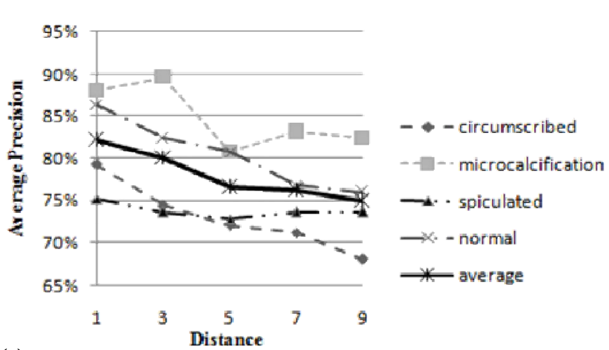
For testing the work two databases has been built under the constraint that only 107 ROIs with circumscribed areas were available. Therefore from every class 80 ROIs for the offline database and 25 ROIs for the query database were selected. So altogether the offline database contains 320 ROIs and the query database 100 ROIs. For every query ROI the number of retrieved images is five so that 500 class affiliations of retrieved images are building the base for every result. The preparation was made through random image selection out of all classified images from the DDSM.

One variable in the feature vector extraction is the neighbourhood system for the GLAM generation. In this work the considered neighbourhood systems are symmetric systems built out of four pixels around the centre with several different distances to the centre. Thereby distance equals one means that the direct neighbours in all four directions are building the neighbourhood system.

For performance measurement the parameter of the average precision is used. As average precision the parameter is denoted that arises when the number ofthe correct retrieved imagesis divided by the number of all retrieved images. Thereby a correct image is one that is from the same class than the query image. In Figure 4 the average precision with different distances is shown separately for the first five retrievedimages as well as for the first retrievedimage. Although looking at both cases the highest average precision over all classes can be foundwith the distance equals one, the coherence between precision and distance for the first retrieved image is marginal whereas for the first five retrieved imagesit is distinct.

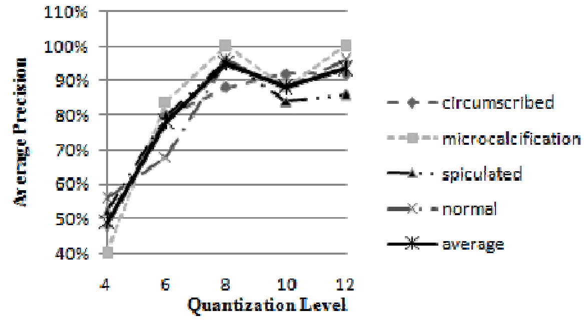
As there is a clearrelationship between quantization and feature size the coherence between allowed numbers of gray levels (quantization level) and average precision is important. As it can be seen in Figure 5 it is obvious that the best result is reached with eight allowed gray levels. The performance decline with less than eight gray levels is so immense that a further reduction of gray levels is not useful despite the smaller feature vector. The results with a quantization level above eight are slightly worse than the best result and additionally the size of the feature vector is bigger.

The best achieved result with an average precision of 82.2% for five retrieved images and 95% for one retrieved image is evident in Figure 4 and Figure 5 as it is reached under the condition of eight allowed gray levels and pixel distance equals one. As the related works always measured the average precision of five retrieved only this result can be compared. The most similar used classification is the classification according to normal, benign and cancer mammograms wherefore

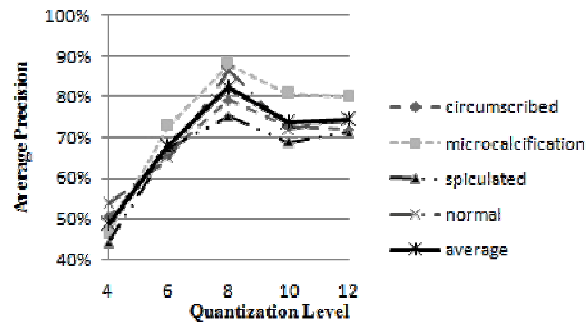


(a)

Figure 4.Average Precision of (a) five and (b) one retrieved images on the condition of eight allowed gray levels and variable pixel distance



(a)



(b)

Figure 5.Average Precision of (a) five and (b) one retrieved images on the condition of variable number of allowed gray levels and pixel distance equals one

the comparison with these works is the most reasonable. The average precision of 70.91% reached with the method of wavelet optimization [4] is clearly outperformed. Also the reached 82.1% under usage of the Dezert-Smarandache Theory [5] are achieved with this work. Therefore the CBIR under the usage of GLAM is a good alternative.

5. CONCLUSION

In this work content based image retrieval (CBIR) system for four types of mammograms from the DDSM database has been implemented under usage of MATLAB®. The necessary texture feature extraction is realized on the basis of GLAM. The functionality of texture extraction using GLAM was demonstrated using average precision whereby better results have been reached than in comparable works with different mammogram classes. Additionally the building of normalized GLAMs is completely independent from shape and size of the ROI which implicates a huge freedom for the user of the CBIR system concerning the considered region. The investigations about the influence of neighbourhood system distance and number of allowed gray levels showed the best result under the conditions of distance equals one and number of allowed gray levels equals eight.

Future works can improve the database creation to have maximal retrieval performance. Furthermore the influence of different and especially asymmetric neighbourhood systems can be investigated and the shapeof the considered ROI can be changed from rectangular to the shape of the actual mark.

To obtain a bigger database either the current mammogram classes has to be rethought or a different database than the DDSM has to be used for retrieves the mammograms which means that most probably a completely new database has to be created from the ground up.

ACKNOWLEDGMENT

The images from IRMA are processed images from the DDSM courtesy of TM Deserno, Dept. Of Medical Informatics, RWTH Aachen, Germany.

REFERENCES

[1] Chia-Hung Wei, Chang-Tsun Li, and Roland Wilson, "A General Framework for Content-Based Medical Image Retrieval with its Applicaton to Mammograms," *Proceeding of SPIE International Symposium on Medical Imaging*, 2005, pp. 134-143.

[2] Remco C. Veltkamp and MirelaTanase, "Content-Based Image Retrieval Systems: A Survey,"*Technical Report UU-CS-2000-34*,2005.

[3] Ranjith M, Balaji R.M., Surjith Kumar M, Dhyaneswaran J, and Baskar A, "Content based Image Retrieval for Medical Image (cerebrum infract) using PCA," *RTCSP Conference Proceeding*, 2009, pp. 124-126.

[4] G. Quellec, M.Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet Optimization for Content-Based Image Retrieval in Medical Databases," *Medical Image Analysis, Volume 14, Issue 2*, 2010, pp. 227-241.

[5] GwénoléQuellec, Mathieu Lamard, Guy Cazuguel, Christian Roux, and BéatriceCochener, "Multimodal medical case retrieval using the Dezert-Smarandache Theory," *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 394-397.

[6] Júlia E. E. de Oliveira, Ana Paula B. Lopes, Guillermo Cámara-Chavez, and Arnaldo de A. Araújo, "MammoSVD: a Content-Based Image Retrieval System Using a Reference Database of Mammographies," *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems*, 2009, pp.1-4.

[7] Júlia E. E. de Oliveira, and Arnaldo de A. Araújo, "MammoSysLesion: a Content-Based Image Retrieval System for Mammographies," *IWSSIP 2010 – 17th International Conference on Systems, Signals and Image Processing*, 2010, pp. 408-411.

[8] JelenaBozek, Mario Mustra, KresimirDelac, and MislavGrgic, "A Survey of Image Processing Algorithms in Digital Mammography," *Recent Advances in Multimedia Signal Processing and Communications*, 2009, pp.631-657.

[9] Xuejie Qin, and Yee-Hong Yang, "Similarity Measure and Learning with Gray Level Aura Matrices (GLAM) for Texture Image Retrieval," *IEEE CVPR*, 2004, pp.326-333.

[10] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W. Philip Kegelmeyer, "The Digital Database for Screening Mammography," *Proceedings of the Fifth International Workshop on Digital Mammography*, M.J. Yaffe, ed., Medical Physics Publishing, 2001, pp.212-218.

[11] Michael Heath, Kevin Bowyer, Daniel Kopans, W. Philip Kegelmeyer, Richard Moore, Kyong Chang, and S. MunishKumaran, "Current status of the Digital Database for Screening Mammography," *Proceedings of the Fourth International Workshop on Digital Mammography*, Kluwer Academic Publishers, 1998, pp. 457-460.