Journal of Engineering Science and Technology
Vol. 10, No.11 (2015) 1453 - 1464
© School of Engineering, Taylor's University

AN INTELLIGENT CONTENT BASED IMAGE RETRIEVAL SYSTEM FOR MAMMOGRAM IMAGE ANALYSIS

K. VAIDEHI*, T. S. SUBASHINI

Department of Computer Science and Engineering, Annamalai University,
Annamalai Nagar, 608002, Tamil Nadu, India
*Corresponding Author: vainakrishna@gmail.com

Abstract

An automated segmentation method which dynamically selects the parenchymal region of interest (ROI) based on the patients breast size is proposed from which, statistical features are derived. SVM classifier is used to model the derived features to classify the breast tissue as dense, glandular and fatty. Then K-nn with different distance metrics namely city-block, Euclidean and Chebchev is used to retrieve the first k similar images closest to the given query image. The proposed method was tested with MIAS database and achieves an average precision of 86.15%. The results reveals that the proposed method could be employed for effective content based mammograms retrieval.

Keywords: Content-based image retrieval, Computer-aided diagnosis, Support vector machine, Statistical descriptors, Breast density.

1. Introduction

Globally, the breast cancer incidence is alarmingly increasing year by year. According to the statistics released by the International Agency for Research on Cancer (IARC), 14.1 million new cancer cases and 8.2 million deaths were reported during 2012 [1]. For detection and diagnosis of breast cancer, radiologist highly depends on mammograms. Mammography helps the radiologist to look for cancer in women having symptoms or no symptoms of breast cancer. Physician recommends screening mammography to the patients who is above 40 years of age. Interpreting the screening mammography is difficult because the level of radiologist experience and image quality affects the screening mammography sensitivity. Normally the mammogram is subjected to double reading by another radiologist. A single reading may result in false negatives and false positives. CAD systems are now a day employed to act like a second radiologist which

Nomenclatures

cb_{12}	Chebychev distance between two feature vectors
ct_{12}	City block distance between two feature vectors
eu_{12}	Euclidean distance between two feature vectors
K	kth element of the feature vcector
n	Total number of elements in the feature vector
N	Total number of pixels in an image
x_{1k}	Feature vector of the query image
x_{2k}	Feature vector of the image in the database
X_{ij}	Pixel intensity at the index location *ij* where *i* represents row and *j* represents column

Greek Symbols

μ	Mean intensity of the image
σ	Standard deviation

Abbreviations

CBIR	Content-Based Image Retrieval
IARC	International Agency for Research on Cancer
SVM	Support Vector Machine

assists in cancer detection and diagnosis [2]. CBIR is an emerging technology which helps in retrieving mammograms similar to the mammogram of the patient under diagnosis. This helps the radiologist to analyse previous diagnostic results of the similar pathologies and help the doctor to arrive at accurate decisions. Day to day enormous amount of images are produced in the medical domain, managing the database, diagnosing and retrieving the same pathology image is a herculean job for the radiologist. CBIR effectively manages the image databases by automatic image indexing and retrieving the visually similar and clinically relevant images which are important for clinical decision making process. The proposed CBIR system based on breast tissue characters (content) is useful for classifying and retrieving similar mammogram images from huge mammogram database and archives. Here content means statistical properties extracted from the images, these are called features, which help for better classification and retrieval of mammograms.

This work is carried in two distinct steps 1) classifying the mammograms based on the type of breast tissue density using SVM as a classifier 2) search and retrieving the top 5 mammograms from the classified mammogram database using KNN. Distance metrics namely Cityblock distance, Euclidean distance, and Chebychev is applied in k-NN algorithm to retrieve the images. This proposed system is tested with mini-MIAS database [3].

The paper is structured as follows: Section 2 gives a survey of the literature done. A methodology of the proposed work is given in the section 3. Brief description of the experiments done and the results obtained is given in the section 4 and section 5 concludes the paper.

2. Literature Review

A survey of the image processing and pattern analysis techniques used by the various researchers in CAD for breast cancer is presented in [4]. Features were estimated and spatial gray level dependency matrices were constructed to characterize the breast tissue [5]. Mammogram retrieval system developed in [6], uses shape, histogram, texture, moments, granulometric and radon features to retrieve mammograms based on breast density patterns. The graph cut segmentation technique is proposed for visualizing the breast anatomical regions in [7]. 2DPCA, PCA and SVD features were applied to retrieve mammogram images based on breast density and lesions [8]. In this work SVM with Gaussian kernel and polynomial kernel is used for classification to evaluate the features. Features related to shape and margin of the mass are extracted from the cancerous regions which were used to retrieve similar mammogram images in [9].

The author in [10] used statistical moments to characterize the breast tissue based on the BIRADS categories. Similar mammogram images were retrieved based on the breast mass is proposed in [11]. SVM and neural network were used in [12] for retrieving calcification clustered images. Relevance feedback is used to guide the retrieval process. The users perception of similarity is predicted from training examples. A CBIR system [13] is developed using singular value decomposition features and histogram. SVM with linear, radial and polynomial kernals are investigated and pattern similarity is computed to separate the four BI-RADS categories. In our previous work [14] statistical moment with SVM classifier is used for mammogram tissue classification.

3. Proposed Method

The density of the breast tissue is highly connected with breast cancer. Women with dense breast have more probability to get breast cancer than women with other fatty or glandular tissue. This necessitates the development of a CBIR system which is based on the tissue type. The major objective of the study to retrieve the mammogram images based on the breast tissue density of the given query mammogram image. The proposed framework is illustrated in Fig. 1.

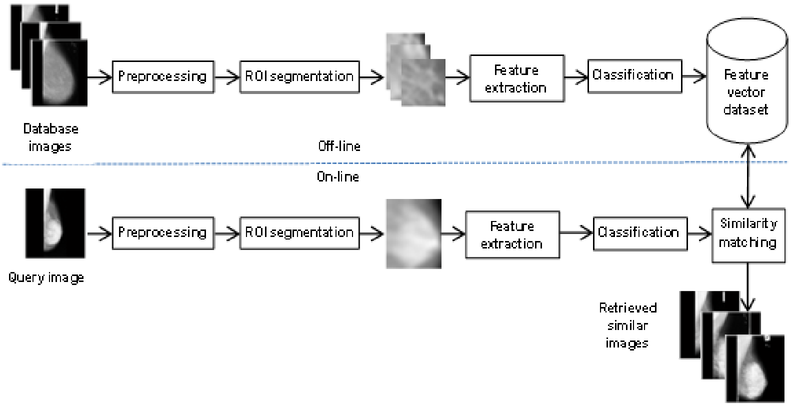

Fig. 1. Block diagram of the proposed work.

In this proposed method region of interest is segmented from the original image, before extracting statistical features. The extracted features are classified using SVM classifier. After classification the classified images and features are stored and retrieved using K-nn with City-block distance, Euclidean distance and Chebychev distance as distance metrics.

3.1. Preprocessing

Preprocessing stage consists of two processes artifacts removal and pectoral muscle removal. High intensity radio opaque artifacts may lead to misclassification and pectoral muscles are the predominant region which may affect the detection of breast density so before actual segmentation is performed the mammogram images are preprocessed for reducing the misclassification rate and used for further processing. Artifacts are removed using our previous work on artifact removal [14]. Straight line method is used for pectoral muscle removal. Median filtering and morphological operations are used for enhancing the image [15]. Artifact and Pectoral muscle removed images are shown in Fig. 2.

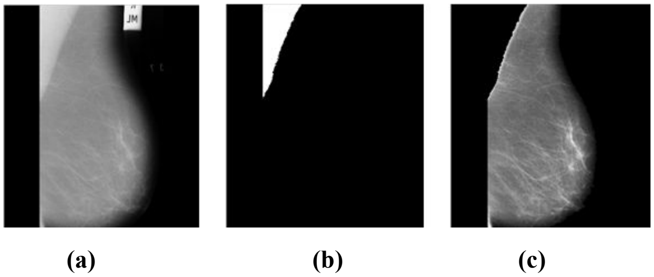

Fig. 2. (a) Original image, (b) Pectoral muscle, (c) Pectoral muscle removed image.

3.2. ROI segmentation

Since most of the mammogram image contains dark background, the parenchymal region is alone segmented by applying the bounding box region property and it is shown in Fig. 3(a). Since most of the upper portion contains only fatty tissue, these may produce false positive during classification and hence these regions are eliminated and the region of interest is obtained as a rectangle using the proposed algorithm. Since, the mammogram size depends upon the patient’s breast size the proposed algorithm dynamically decides the size of the ROI rectangle depending on the size of the parenchymal region.

The proposed procedure to obtain the region of interest is as follows:

- The region of interest lies at the bottom of the image, Fig. 3(b). The center point of the ROI rectangle is at 2/3 of the length and ½ of the width of the image.
- ½ of the width is taken as the length of the ROI rectangle and 8/20 of the width is taken as the breadth of the ROI rectangle. These values are found out empirically. Now we get the region of interest which is shown in Fig. 3(c).

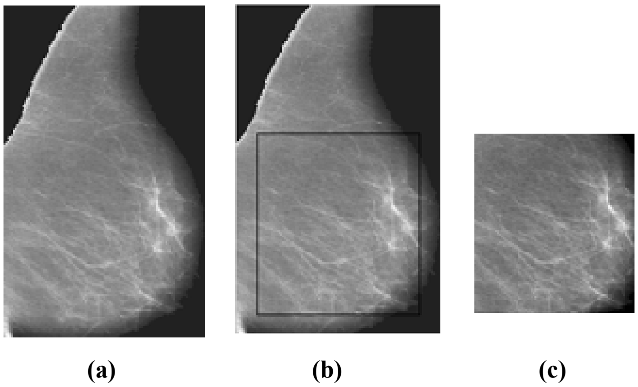

Fig. 3. (a) Breast parenchymal portion alone, (b) Rectangle portion will be segmented from the breast parenchyma, (c) Segmented region of interest.

3.3. Statistical feature extraction

CBIR refers to the retrieval of similar images from the image database, using measures of information derived from the images themselves [16]. Here information refers to some relevant descriptors which are representative of the whole image. Statistical features give more significant information in pattern recognition area and in this work statistical features are extracted. Statistical feature extraction methods characterize texture by the statistical distribution of the image gray level intensity [17, 18]. Statistical methods can be classified into first order (mean), second order (variance) and higher order (skewness, kurtosis) statistics. The first order statistics does not consider the spatial relationship with neighboring pixels and the other higher order statistics evaluates the relationship with the neighboring pixels is estimating the properties of the current pixel [18].

In our work, 11 statistical features such as mean, standard deviation, smoothness, skewness, uniformity, kurtosis, average histogram, median, mode, modified standard deviation, modified skewness were derived from the segmented region of interest [14]. The SVM was trained with various combination of these extracted features at the best performance was achieved using four features namely mean, standard deviation, skewness and kurtosis. Table 1 gives the descriptors used in training and testing the SVM. The accuracy of 92.18% was achieved and it reduced seven more features were included and the optimal set of four features were arrived.

Table 1. Statistical descriptors used in this work.

Descriptors	Mathematical Expression
Mean [14]	$\mu = \frac{\sum_{ij} X_{ij}}{N}$
Standard deviation [14]	$\sigma = \sqrt{\frac{\sum_{ij}\left(X_{ij} - \mu\right)^2}{N}}$
Skewness [14]	$\frac{\sum_{ij}\left(X_{ij} - \mu\right)^3}{N\sigma^3}$
Kurtosis [14]	$\frac{\sum_{ij}\left(X_{ij} - \mu\right)^4}{(N-1)\sigma^4}$

3.4. Classification and Retrieval

SVM classifier is a simplest supervised classifier [19]. Based on the literature [14, 20, 21] for texture based classification SVM is found to give better performance than other classifiers. So it is proposed to be SVM for classification in our work. SVM with polynomial kernel is used for classifying the mammograms into three different tissue classes namely dense, glandular and fatty. These classified images and its corresponding feature vector are stored in a database. K-nn with Euclidean distance metric, city block distance metric and Chebychev distance metric is used for similarity matching and top -10 similar images are chosen from the classified images. Table 2 gives the expressions of distance metrics. Figure 4 shows the flowchart of the SVM training and testing phases.

Table 2. Distance metrics used in this work.

Distance metrics	Mathematical Expression
Euclidean	$eu_{12} = \sqrt{\sum_{k=1}^{n}\left(x_{1k} - x_{2k}\right)^2}$
City block	$ct_{12} = \sum_{k=1}^{n}\lvert x_{1k} - x_{2k}\rvert$
Chebychev	$cb_{12} = \max_{i}(\lvert x_{1_i} - x_{2_i}\rvert)$

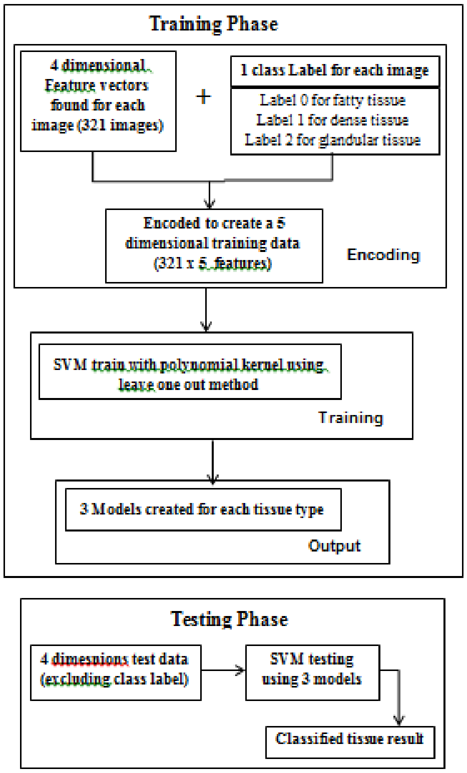
Training Phase
4 dimensional Feature vectors found for each image (321 images)
+
1 class Label for each image
Label 0 for fatty tissue
Label 1 for dense tissue
Label 2 for glandular tissue
Encoded to create a 5 dimensional training data (321 x 5 features)
Encoding
SVM train with polynomial kernel using leave one out method
Training
3 Models created for each tissue type
Output
Testing Phase
4 dimesnions test data (excluding class label)
SVM testing using 3 models
Classified tissue result

Fig. 4. Flowchart of SVM training and testing phases.

3.5. Evaluation measures

Precision-recall performance metric [22, 23] is used for evaluating the performance of the retrieval system. Precision is the ratio of the number of relevant images retrieved to the number of total images retrieved. Recall is the number of relevant images retrieved over the total number of relevant images available in the database. The precision recall curve measures the effectiveness of the CBIR system for retrieving most similar images. And the Retrieval performance of similarity measures is also used to measure the performance evaluation of CBIR system.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Number of images retrieved}}\text{x}100 \tag{1}$$

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Number of relevant images in the database}}\text{x}100 \tag{2}$$

4. Experimental results

The first four order statistical moments namely mean, standard deviation, skewness and kurtosis are computed from the ROI. 322 mammograms from the Mini-MIAS database have taken up for this study. SVM classifier determines the tissue class based on the feature vector created. Leave one out procedure has been adopted in testing the performance of the SVM classifier.

The SVM is trained in multiclass mode and achieves an overall accuracy of 92.18%. Table 3 shows the classification accuracy of SVM using polynomial kernel. Table 4 shows the comparison of our proposed work with previous works on classification. All the work in the Table 3 was tested with MIAS dataset images.

4.1. Retrieval results

The classified images and its corresponding features are stored in a database. K-nn with city block distance, Euclidean distance metric and Chebychev distance metric is used for similarity matching and top -10 similar images are retrieved from the classified images. Figures 5, 6 and 7 show the top 5 images retrieved for the given dense, glandular and fatty query mammograms. The images retrieved are ranked by degree of similarity in accordance to the query image.

4.2. Performance analysis

The performance of the proposed system is evaluated using standard performance metrics namely precision and recall. 120 query images were randomly selected from the 322 images of MIAS database. Precision and recall rates calculated for 12, 24, 36, 48, 60, 72, 84, 96, 108 and 120 retrieved images are used to plot the PR graph which is shown in Fig. 8. From the graph it could be seen that the first point (i.e.) the left top most point represents the highest precision which indicates that the first retrieved image is same as the query image.

The bar chart in Fig. 9 shows the retrieval performance of top 10 retrieved images using various distance measures namely cityblock, Euclidean and Chebychev distance metrics. Average precision of top 10 images using City block distance, Euclidean distance and Chebychev distance is 91.28%, 93.71% and 94.20% respectively. Average time for retrieving top 10 images using City block distance, Euclidean distance and Chebychev distance is 0.40s, 0.43s and 0.37s respectively. Chebychev distance achieves higher precision than Euclidean and Cityblock distances. The overall precision of City block distance, Euclidean distance and Chebychev distance is 77.91%, 84.65% and 86.15% respectively.

Table 3. Results of breast tissue classification using SVM classifier.

Classifier	SVM		
Tissue density	Density	Glandular	Fatty
Correct classification	105	93	99
Missed classification	7	12	7
Accuracy in(%)	93.75	89.42	93.39

Table 4. Comparison between our proposed work and previous work for classification.

No.of images	Features	Classifier	Accuracy	Reference
43	Statistical features (Whole breast)	SVM	95.44%	T.S.Subashini et al. [14] 2010
322	Fractal features (Whole breast)	SVM	85.7%	S.D.Tzikopoulos et al. [22] 2011
186	SIFT, LBP, texton, histogram (ROI)	SVM	93.54%	G.Liasis et al. [23] 2012
322	GLCM, Statistical, Histogram (ROI)	K-nn	82.5%	M.Mario et al., [24] 2012
322	Statistical moments (ROI)	SVM	92.18%	Proposed Method

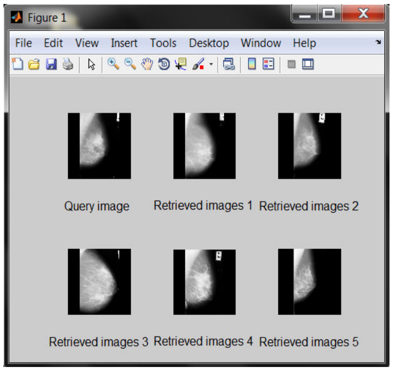

Fig. 5. Retrieval of dense images.

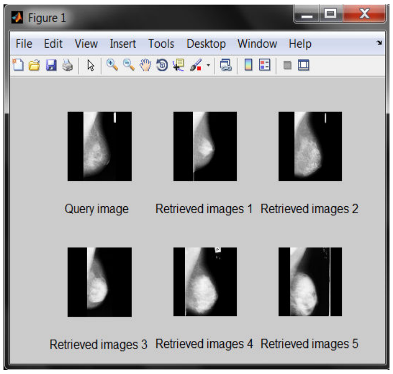

Fig. 6. Retrieval of glandular images.

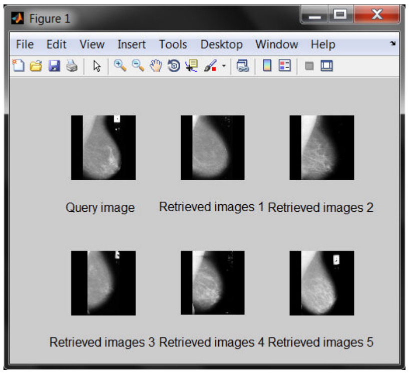

Fig.7. Retrieval of fatty images.

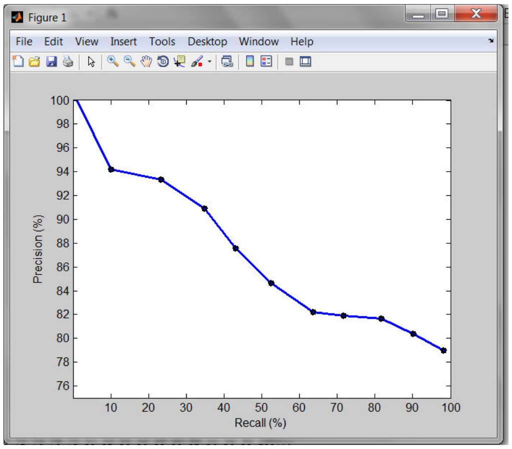

Fig. 8. Precision-Recall graph.

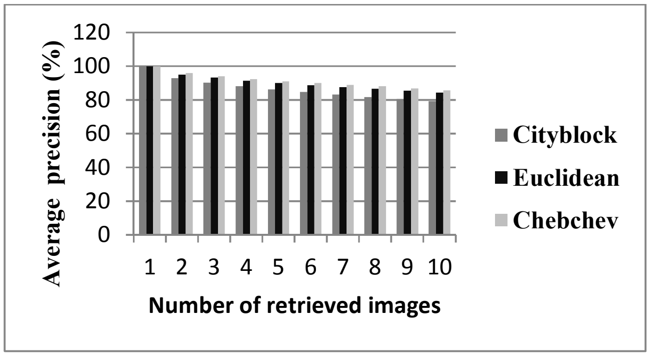

Fig. 9. Retrieval performance of top k retrieved images (k=1 to 10) that actually match the query.

5. Conclusion

The proposed work was carried out using MATLAB R2012a (Version 7.14). In this work, the rectangular region from the bottom region of the mammogram which characterizes the breast tissue effectively is segmented automatically. The size of the rectangular region segmented varies with respect to the patient's breast size. This is the region of interest from which statistical moments are derived. The derived features are modeled with SVM to classify the breast tissue into fatty, dense or glandular breast. The feature vectors along with the respective image are stored in the database for retrieval purpose. Images similar to the given query image are retrieved using K-nn algorithm with City-block distance, Euclidean distance and Chebychev distance as distance metrics with Chebychev outperforming others with the overall precision of 86.15%. And this work can be used in the processing chain to adapt parameters for classification and retrieval of breast lesions.

Acknowledgements

The authors would like to thank Dr.M.K.Sivakkolunthu, Professor of Radiology, Raja Muthiah Medical College Hospital, Annamalai Nagar for his valuable help and comments in carrying out this work. The work has been done under University Grants Commission (UGC) Major Research Project. The financial support of UGC is greatly acknowledged with appreciation.

References

1. Gaudin, N.. (2013). The International Agency for Research on Cancer. World Health Organisation. Lyon/Geneva. 12th December 2013. Pr223_E.pdf.
2. Rangayyan, R.M.; Ayres, F.J.; and Leo Desautels, J.E. (2007). A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, 344(3), 312-348.
3. Suckling, J.; Parker, J.; Dance, D. et al.,. (1994).The Mammogram Image Analysis Society Digital Mammogram Database. *Exerpta Medica,* International Congress Series, 1069, 375-378.

4. Rangayyan, R.M. (2005). *Biomedical Image Analysis*, CRC press LLC.
5. Bovis, K.; and Singh, S. (2002). Classification of mammographic breast density using combined classifier paradigm. *Proceedings on Medical Image Understanding and Analysis*.
6. Kinoshita, S.K.; Azevado-Marques, P.; Pereira,R.; Rodrigues, J.; Rangayyan, R. (2007) Content-based Retrieval of Mammograms Using Visual Features Related to Breast Density Patterns, *Journal of Digital Imaging*, 20(2), 172-190.
7. Nafiza Saidin; Harsa Amylia Mat Sakim; Umi Kalthum Ngah; Ibrahim Lutfi Shuaib. (2013). Computer Aided Detection of Breast Density and Mass, and Visualization of other Breast Anatomical Regions on Mammograms using Graph Cuts. *Computational and Mathematical Methods in Medicine*, http://dx.do i.org/10.1155/2013/205384.
8. Oliver, J.E.E.; Araújo, A.A.; Deserno, T.M.; (2010) MammoSysLesion: a Content-Based Image Retrieval System for Mammographies, *IWSSIP 2010-17th International Conference on Systems, Signals and Image Processing*.
9. Wei, C.-H.; Chen, S.Y.; Liu, X. (2012). Mammogram retrieval on similar mass lesions. *Computer Methods and Programs in Biomedicine*, 234-248.
10. Sheshadri, H.S.; Kandaswamy, A. (2006). A Breast tissue classification using statistical feature extraction of mammograms, *Med Imag Inform Sci*, 23, 105-107.
11. Muramatsu, C.; Li, Q.; Suzuki, K.; Schmidt, R.A.; Shiraishi, J.; Newstead, GM.; Doi, K. (2005). Investigation of psychophysical measure for evaluation of similar images for mammographic masses: preliminary results. *Med Phys*, 32, 2295–2304.
12. Issam El-Naqa; YongyiYnag; Nikolas, P.Galastsanos; Robert, M.Nishikawa; Miles, N.Wernick. (2004). A Similarity Learning Approach to Content-Based Image Retrieval: Application to Digital Mammography. *IEEE Transactions on Medical Imaging*, 23(10), 1233.
13. Oliveira, J. E.E.D.; Araujo, A.D.A.; Deserno, T.M.; (2011). Content-based image retrieval applied to BI-RADS tissue classification in screening mammography. *World J Radiol*, 3(1) 24-31.
14. Subashini, T.S.; Ramalingam, V.; Palanivel. S. (2010). Automated assessment of breast tissue density in digital mammograms. *Computer Vision and Image Understanding*, 114(1), 33-43.
15. Vaidehi, K.; Subashini, T.S. (2013). Automatic Identification and elimination of pectoral muscle in digital mammograms. *International Journal of Computer Applications*, 75(14), 15-18.
16. Marques, P.A.; Rangayyan, R.M. (2013). *Content –based Retrieval of Medical images*, Morgan & Claypool Publishers.
17. Chandy, D.; Abraham, J.; Stanly Johnson; S. Easter Selvan. (2013). Texture feature extraction using gray level statistical matrix for content-based mammogram retrieval. *Multimedia Tools and Applications*, 1-14.
18. Srinivasan, G.N.; and Shobha, G. (2008). Statistical texture analysis, *proceedings of world academy of science, engg& tech*, 36.
19. Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, New York.

20. Tzikopoulos, S.D.; Mavroforakis, M.E.; Georgiou, H.V.; Dimitropoulos; Theodoridies, N. S. (2011). A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry, *Computer Methods and Programs in Biomedicine*, I02, 47-63.
21. Liasis, G.; Pattichis, C.; Petroudi, S. (2012). Combination of different texture features for mammographic breast density classification. *IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, 732-737.
22. Muller, H.; Muller, W.; Squire, DM.; Marchand-Maillet, S.; Pun, T. (2005). Performance evaluation in content based image retrieval: overview and proposals. *Pattern Recognition Letters*, 22(5), 593-601.
23. Tourassi, G.; Harrawood, B.; Singh, S.; Lo, J; Floyd, C. (2007). Evaluation of information theoretic similarity measure for content-based retrieval and detection of masses in mammograms. *Medical Physics*, 34, 140-150.
24. Mario, M.; MislavGrgić; KrešimirDelač. (2012). Breast Density Classification Using Multiple Feature Selection. *AUTOMATIKA*, 53(4), 362-372.