33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

A WAVELET-PACKET-BASED APPROACH FOR BREAST CANCER CLASSIFICATION

Meysam Torabi[1], Seiied-Mohammad-Javad Razavian[2], Reza Vaziri[3], and Bijan Vosoughi-Vahdat[3]

ABSTRACT

In this paper, a new approach for non-invasive diagnosis of breast diseases is tested on the region of the breast without undue influence from the background and medically unnecessary parts of the images. We applied Wavelet packet analysis on the two-dimensional histogram matrices of a large number of breast images to generate the filter banks, namely sub-images. Each of 1250 resulting sub-images are used for computation of 32 two-dimensional histogram matrices. Then informative statistical features (e.g. skewness and kurtosis) are extracted from each matrix. The independent features, using 5-fold cross-validation protocol, are considered as the input sets of supervised classification. We observed that the proposed method improves the detection accuracy of Architectural Distortion disease compared to previous works and also is very effective for diagnosis of Spiculated Mass and MISC diseases.

***Index Terms*— Non-invasive diagnosis, statistical feature extraction, supervised classification, Wavelet packet analysis, breast diseases.**

I. INTRODUCTION

Digital image processing techniques have been applied to digitized images of breast during the past 10 to 20 years (eg. [1], [2] and many others) for various purposes e.g. image quality improvement, mammographic feature enhancement and malignant sign identification/analysis [3]. (Fig. 1)

Computer Aided Diagnosis (CAD) systems provide complementary diagnostic information to help physicians, since early detection of breast cancer increases the survival rate and also increases the treatment options [4].

Although breast cancer is one of the most important diseases in women, the certain diagnosis of this disease has not been reported yet [2]. Therefore, the successful development of CAD systems will be invaluable, if they can detect breast abnormalities in the first steps of the disease.

Women who suffer from breast diseases usually have two concerns. The first one is that which kind of abnormalities is occurred, and the second concern is about how serious the disease is, meaning whether it is benign or malignant.

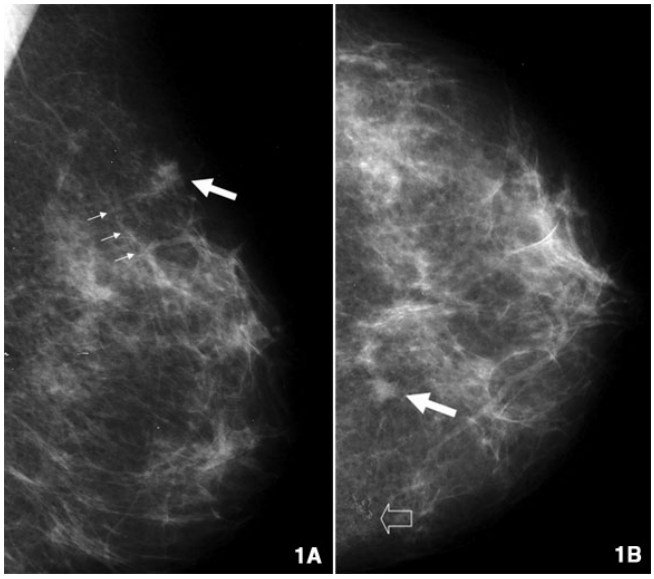

Fig. 1. A 64-year-old asymptomatic female presents for screening and subsequent diagnosis: 1A) Left mediolateral oblique demonstrates a mass in the upper breast (large arrow) while the smaller arrows denote an artery for anatomic reference, 1B) Craniocaudal view shows that the mass is in the inner left breast (solid arrow) dust artifact is noted medial to this finding (open arrow); (taken from [5]).

In our previous work [6], texture analysis was used to diagnose breast cancer and categorize four different breast diseases, however satisfiable results in detection of Architectural Distortion disease was not achieved. In that work, we approached the problem by applying texture analysis directly on images without any particular preprocessing. The contribution of the present paper is applying Wavelet Packet Analysis (WPA) on images as the preprocessing stage before extraction of textural features, in order to improve recognition rate of Architectural Distortion disease, and also to enhance detection of Ill-defined Mass (or MISC) and Spiculated Mass, while the recognition rates of other three previously studied diseases (i.e. Circumscribed Mass, Asymmetry, Calcification diseases) are still appropriate.

In order to classify breast images, four subbands (namely

This work is entirely done at Sharif University of Technology, Tehran, Iran.

[1]M. Torabi is a Graduate student in UC Berkeley, CA, USA. torabi@berkeley.edu

[2]Seiied-Mohammad-Javad Razavian is graduated from Computer Science Department of Sharif University of Technology, Tehran, Iran.

[3]Reza Vaziri and Bijan Vosoughi-Vahdat are with The School of Electrical Engineering of Sharif University of Technology, Tehran, Iran.

sub-images) are computed for each image by WPA. Then, texture analysis which has been frequently used in breast studies (e.g. [6], [7], [8], [9] and [10]) is applied on each sub-image. After reduction of features, two trained neural networks are used as the supervised nonlinear classifiers to categorize defected breasts into the six diseases and also to detect the malignancy.

The dataset used in this work is provided by Mammography Image Analysis Society (MIAS) composed by 322 left and right breast images [11].

In section II, related works are reviewed. The section III describes the proposed method, while the next section will illustrate our classification results. Finally, the last sections will cover conclusions, summary and future works.

II. RELATED WORK

Breast cancer persists to be the top threat to women's health. Thus, there are many works in literature on how to investigate breast cancer diseases. Among them, image processing methods are extensively used, because they are fully non-invasive.

Chandra *et al.* [12] classified and clustered medical image data including normal and malignant breasts, using self-organizing neural networks with quadratic neural type junctions, and observed that the self-organizing maps classify breast images effectively.

Based on edge-sharpness, shape and texture of images, Nandi *et al.* [13] computed 22 features from 57 breasts. They adapted and applied genetic programming and used three statistical measures Student's t-test, Kolmogorov-Smirnov test and Kullback-Leibler divergence. The results showed an appropriate accuracy.

In the field of ultrasound image analysis, Rodrigues *et al.* [10] applied an initial image segmentation by maximization of the information between the foreground and background using a new kind of entropy i.e. non-extensive entropy, that was generalized as Shannon entropy. They combined the q-entropy, a level set formulation and a support vector machine framework to classify benign/malignant breasts.

Since the classification of breasts is the most important stage of breast cancer studies, different kinds of classifiers are recommended in literature. In [2], [10], [14] support vector machines as learning framework are performed as classifier, while in [6], [15], [16] neural network is used. Also Essafi *et al.* [17] developed Principal Component Analysis (PCA) to classify abnormal breasts.

Researchers have used different image datasets to validate their methods (e.g. MIAS [18], [19], [20], WBCD [15], DDSM [19], [20] and CBCP [21]). Moreover, some references used two datasets. Oliver *et al.* [19] exploited a kindred topic in computer vision for face detection/classification and adapted it to breast mass detection and then tested it by two different datasets MIAS and DDSM. They reported initial results to demonstrate the feasibility of using such approaches. Also, another work accomplished by Bosch *et al.* [20] can be mentioned in which DDSM and MIAS were both examined to prove generality of the method.

However, obviously, it is preferred to develop a comprehensive CAD system to recognize benign/malignant cases, and also concurrently classify images into different kinds of abnormalities as pursued in the present paper.

III. METHOD

As mentioned, in this paper WPA is used as a preprocessing stage before extraction of textural features. Before using WPA, marginal parts of images which do not have medical data are automatically removed based on image histogram analysis. Then, image background is omitted by setting an adaptive gray-level threshold.

In the next step, two-dimensional histogram, which is also in some literature called Gray-Level Co-occurrence Matrix (GLCM), is calculated for each of sub-images and then is normalized before extraction of features.

III-A. Wavelet packet analysis

The WPA is a generalization of Wavelet decomposition which offers a richer range of possibilities for signal analysis. In Wavelet analysis, a signal is split into an approximation and a detail. The approximation is then itself split into a second-level approximation and detail, and the process is repeated. For an n-level decomposition, there are $n+1$ possible ways to decompose or encode a digital image [22].

In WPA, the details as well as the approximations would be split.

III-B. Calculation of gray-level co-occurrence matrix

Obviously, features resulted from first-order histogram provide information that are just related to the gray-level distribution and hence do not give any information about pixels' position of various gray-levels within the image.

Hence, GLCM is used to overcome this problem which is a common approach for texture analysis. In this section, GLCM will be calculated from the sub-images which were previously obtained by WPA.

In order to calculate GLCM's, d is defined as the distance among the pixel numbers ($d = 1, 2$) and θ is defined as the distance direction along some angle. With n distances and m directions, $n \times m$ numbers of GLCM's can be computed. In this regard, the element $H(g_i, g_j)$ of GLCM is defined by

$$H(g_i, g_j) = \frac{\text{Number of pairs of pixels at distance } d \text{ and angle } \theta \text{ with value } (g_i, g_j)}{\text{Total number of possible pairs}}. \quad (1)$$

in which the directions and distances are defined as follows

$$H(G(p,q) = g_i, G(p \pm d, q) = g_j), \quad \theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$$

where g_i and g_j are the quantized gray-levels and also address the element $H(g_i, g_j)$ in GLCM, and p and q address the element $G(p, q)$ in the gray-level matrix.

III-C. Definition of statistical features

After calculation of GLCM's, appropriate features can be generated to somehow quantify texture properties such as smoothness or regularity, homogeneous or inhomogeneous, and so forth.

Actually, these features will emerge by exploiting space relations underlying the gray-level distribution. In this way, some particular statistic features as textural features are calculated from each GLCM.

Some of textural features have a direct physical interpretation with respect to the human texture, such as coarseness and smoothness while others do not possess such a property, but still encode information related to texture.

One of the features is "contrast", as a measure of image contrast indicating the local gray-level variations in GLCM. Therefore contrast takes high values for image of high contrast, calculated as follows

$$Cnt = \sum_{n=0}^{N-1} n^2 \left\{ \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} H(i,j) \right\} \quad |i-j| = n \tag{2}$$

where N is the number of quantized gray-levels.

A higher n causes more accuracy and also more time needed to calculate the contrast. The value of n can be manipulated to find an optimum contrast.

Another feature is the "angular second moment" or "uniformity" providing sum of squared elements in GLCM

$$ASM = \sum_{i=0}^{N_g} \sum_{j=0}^{N_g} (H(i,j))^2 \tag{3}$$

which is also a measure of the smoothness of image. The less smooth the region is the more uniformity distributed $H(i,j)$ and the lower the uniformity [22].

The last feature is "correlation" which measures the joint probability occurrence of the specified pixel pairs

$$Crl = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i \times j \times H(i,j) - \mu_x \times \mu_y}{\sigma_x \sigma_y} \tag{4}$$

where μ_x and μ_y are the first moments, and also σ_x and σ_y are the second moments in direction of x and y, respectively.

In addition to the above-mentioned features, also mean, standard deviation, skewness, kurtosis, minimum and maximum gray-levels of sub-images are calculated to totally generate 38 features for each sub-image.

III-D. Supervised nonlinear classification

The main part of this work is classification of breast images which should be accomplished by a powerful discriminator.

Due to number of the features, there is probably some dependency between them, because sub-images are, in the general view, similar to each other and hence some of eigenvalues which belong to the feature covariance matrices may be nearly zero. In this regard, PCA is used to reduce number of the features.

In order to appropriately classify the images, supervised classifiers, i.e. two neural networks, are designed, tuned and utilized. The neural network 'A' splits the dataset into benign/malignant breasts, and concurrently neural network 'B' categorizes them into the six groups of the breast diseases.

The input of neural network 'A' and 'B' contains 26 elements (i.e. 26 features) generated as the output of PCA.

IV. RESULTS

In order to generate more reliable diagnosis results, the 5-fold cross validation is used in which images in each class are categorized in five subsets, four of which are considered as training set while the fifth subset is reserved for test set. This process is repeated 5 times with each of the 5 subsets used exactly once as the validation data. Then, by averaging on the all 5 results, the final result will be obtained, as presented in Table I. In Table II, the recognition rate of benign/malignant cases is reported.

Table I. The classification rates of the six breast diseases

Disease	CIRC	MISC	ASYM	ARCH	SPIC	CALC
Rate	64%	87%	79%	74%	79%	71%

Table II. Recognition rate of benign/malignant cases

Issue	Benign/Malignant
Rate	76%

V. SUMMARY AND CONCLUSION

In the present paper, the six breast diseases are examined using a novel approach with focusing on Architectural Distortion, Spiculated Mass and MISC diseases.

The results show that Wavelet packet decomposition improves the recognition rate of Architectural Distortion disease noticeably and also is appropriate for detection of two important breast diseases i.e. Spiculated Mass and MISC diseases, while the recognition rate of Circumscribed Mass, Asymmetry and Calcification diseases in addition to benign/malignant detection are still suitable.

In this study, the features of the frequency domain were investigated. The features actually were extracted from sub-images obtained from Wavelet packet decomposition. Then, by two trained supervised nonlinear classifiers, the independent features were analyzed to classify abnormal breast images into the six kinds of breast abnormalities and benign/malignant cases. The classifiers were tested by 5-fold cross-validation protocol to enhance validation of the classification.

In calculation of GLCM's, we limited distance-direction to 1 and 2 to increase the recognition response speed.

VI. FUTURE WORK

In this work, after the extraction of the features, redundant features were eliminated to obtain a reasonable number of features for classification. However, this approach is not time-efficient and hence now we are developing a new method to generate initially-independent features without any need to reduce number of the features.

VII. REFERENCES

[1] M. Torabi, H. Moradzadeh, R. Vaziri, R. D. Ardekani, and E. Fatemizadeh, "Multiple sclerosis diagnosis based on analysis of subbands of 2-d Wavelet transform," in *Proc. ACS/IEEE Int. Conf. Comp. Sys. and App.*, May 2007, pp. 717–721.

[2] E. Behnamghader, R. D. Ardekani, and M. Torabi, "Another approach to detection of abnormalities in MR-images using support vector machines," in 5^{th} *Intl. Symp. on Image and Signal Processing and Analysis*, Sep 2007, pp. 98–101.

[3] K. Bovis and S. Singh, "Classification of mammographic breast density using a combined classifier paradigm," in Proc. 4^{th} *Int. Workshop on Digital Mammography*, 2002, pp. 177–180.

[4] R. Panchal and B. Verma, "Characterization of breast abnormality patterns in digital mammograms using auto-associator neural network," in *Neural Information Processing*, vol. 4234. Springer Berlin / Heidelberg, 2006, pp. 127–136.

[5] E. S. Burnside, A. Trentham-Dietz, F. Kelcz, and J. Collins, "An example of breast cancer regression on imaging," in *Radiology Case Rreports*, vol. 1, no. 2, 2006.

[6] R. D. Ardekani, M. Torabi, and E. Fatemizadeh, "Breast cancer diagnosis and classification in MR-images using multi-stage classifier," in Proc. *Intl. Conf. Bio. Eng. and Pharm. Eng.*, Dec 2006, pp. 70–83.

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," in *IEEE Trans Sys. Man and Cyber*, vol. SMC-3, no. 6, Nov 1973, pp. 610–621.

[8] M. Torabi, H. Moradzadeh, R. Vaziri, S. M. J. Razavian, R. D. Ardekani, M. Rahmandoost, A. Talimi, and E. Fatemizadeh, "Development of alzheimer's disease recognition using automated analysis of statistic parameters based on frequency characteristics of medical images," in *Proc. IEEE Int. Conf. Signal Proc. and Comm.*, Nov 2007.

[9] K. Bovis and S. Singh, "Detection of masses in mammograms using texture features," in Proc. 15^{th} *Intl. Conf. Patt. Recog.*, 2000, pp. 2267–2270.

[10] P. S. Rodrigues, R. F. Chang, and J. S. Suri, "Non-extensive entropy for cad systems of breast cancer images," in Proc. 19^{th} *Brazilian Symp. Comp. Graph. and image Proc.*, Oct 2006, pp. 121–128.

[11] The dataset of Mammography Image Analysis Society (MIAS). [Online]. Available: http://peipa.essex.ac.uk/ipa/pix/mias.

[12] B. Chandra, S. Nath, and A. Malhothra, "Classification and clustering of breast cancer images," in *Intl. Joint Conf. on Neural Networks*, Jul 2006, pp. 3843–3847.

[13] R. J. Nandi, A. K. Nandi, R. Rangayyan, and D. Scutt, "Genetic programming and feature selection for classification of breast masses in mammograms," in 28^{th} *Annu. Intl. Conf. of the IEEE Engineering in Medicine and Biology Society, EMBS '06*. New York, USA: IEEE, Aug 2006, pp. 3021–3024.

[14] S. Shou-kui, W. Xiao-feng, and S. Xi-jing, "Discrimination methods for the classification of breast cancer diagnosis," in Proc. *Intl. Conf. Neural Net. and Brain*, Oct 2006, pp. 259–261.

[15] V. Bevilacqua, G. Mastronardi, F. Menolascina, P. Pannarale, and A. Pedone, "A novel multi-objective genetic algorithm approach to artificial neural network topology optimisation: The breast cancer classification problem," in Proc. *Intl. Joint Conf. Neural Net.*, Jul 2006, pp. 1958–1965.

[16] C. M. Kocur, S. K. Rogers, L. R. Myers, T. Burns, M. Kabrisky, J. W. Hoffmeister, K. W. Bauer, , and J. M. Steppe, "Using neural networks to select Wavelet features for breast cancer diagnosis," in Proc. *IEEE Eng. in Med. and Bio. Magazine*, May/Jun 1996, pp. 95–102.

[17] S. Essafi, R. Doughri, S. M'Hiri, K. B. Romadhane, and F. Ghorbel, "Segmentation and classification of breast cancer cells in histological images," in Proc. 2^{nd} *Conf. Info. and Comm. Tech.*, Apr 2006, pp. 1097–1102.

[18] H. Mirzaalian, M. Ahmadzadeh, and F. Kolahdoozan, "Breast contour detection on digital mammogram," in *Info. and Comm. Tech.*, Apr 2006, pp. 1804–1808.

[19] A. Oliver and J. M. and R. Marti and A. Bosch and J. Freixenet, "A new approach to the classification of mammographic masses and normal breast tissueancer cells in histological images," in Proc. *IEEE* 18^{th} *Int. Conf. Pat. Recog.*, Sep 2006, pp. 707–710.

[20] A. Bosch, X. Munoz, A. Oliver, and J. S. Marti, "Modeling and classifying breast tissue density in mammograms," in Proc. *IEEE Comp. Soc. Conf. Comp. Vision and Patt. Recog.*, 2006, pp. 1552–1558.

[21] S. Maskery, Y. Zhang, H. Hu, C. Shriver, J. Hooke, and M. Liebman, "Caffeine intake, race, and risk of invasive breast cancer lessons learned from data mining a clinical database," in Proc. 19^{th} *IEEE Symp. Computer-Based Med. Sys.*, 2006, pp. 714–718.

[22] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Orlando, FL, USA: Academic Press, May 2006.