BREAST CANCER DIAGNOSIS USING IMAGE RETRIEVAL FOR DIFFERENT ULTRASONIC SYSTEMS*

Yu-Len Huang[†], *Dar-Ren Chen*[‡] *and Ya-Kuang Liu*[†]

[†]Department of Computer Science and Information Engineering Tunghai University, Taichung, Taiwan
[‡]Department of General Surgery China Medical College & Hospital, Taichung, Taiwan

ABSTRACT

This paper employs the image retrieval technique to classify breast tumors as benign or malignant lesions. We evaluated 600 ultrasound (US) images of pathologically proven solid breast nodules including 230 malignant and 370 benign tumors. The US images were acquired from four different ultrasound systems. Firstly, the physician located regions-of-interest (ROI) of ultrasound images. The textual features from ROI sub-image are utilized to classify breast tumors. The principal component analysis (PCA) is used to reduce the dimension of textual feature vector and then the image retrieval technique was utilized to differentiate between benign and malignant tumors. Historical cases can be directly added into the database and training of the diagnosis system again is not needed. The accuracy of the proposed computer-aided diagnosis (CAD) system was 91.2%, the sensitivity was 97.0% and the specificity was 87.6%. This system differentiates solid breast nodules with a relatively high accuracy in the different US systems and helps inexperienced operators avoid misdiagnosis.

1. INTRODUCTION

Early diagnosis and treatment of breast cancers are the most useful way to increase the cure rate [1]. Accurate and reliable diagnostic procedure is important in early diagnosis. Mammography and ultrasonography are the most frequently used tools for early diagnosis of breast tumors. The examinations allowed physicians to differentiate benign breast tumors from malignant ones. Generally, breast ultrasound image played a role as an auxiliary to mammography. However, the ultrasound (US) examination is more convenient and safer than mammography in daily clinical practice. In 1995, Stavros et al. indicated the US technique is helpful to identify breast cancer more accurate [2]. However, if some physicians lacked enough clinical experiences might make a wrong diagnosis from breast US images. In order to avoid unnecessary biopsy and improve the accurate of diagnosis, computer-aided diagnosis (CAD) system can be a second beneficial support for physician to diagnose.

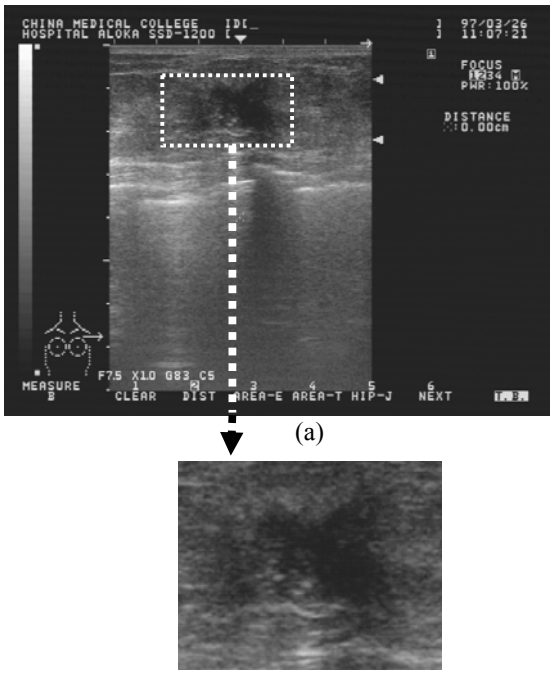
Chen et al. [3-6] used a 2-D auto-covariance and the neural network classifier to differentiate between benign and malignant tumors in breast US images. The variation of texture in US image is a practical feature to identify benign and malignant tumours [5,7]. However, the NN learning procedure [8] is a very time-consuming and the results usually depend on initial parameter setting. Hence, this paper performed the image retrieval scheme by using the auto-covariance coefficients [9] to distinguish malignant from benign masses in the breast US image. The textual feature vector produced by the 2-D auto-covariance matrix is always in a large dimension. The feature vector may carry off an ineffective retrieving result. The proposed CAD system utilizes the principal component analysis (PCA) [10,11] to reduce the dimension of the feature vector. The PCA has been applied in content-based image retrieval [11] and image compression [12] applications. The textural feature vector can be transformed into a lower dimension by using the PCA technique. The transformed vector is used as the new textural feature to select the images from database by a similarity measure of minimum Euclidean distance. The retrieved images apply as the reference materials to identify benign and malignant lesions in the US image [13-17].

2. TEXTUAL ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

In this study, a physician first extracted the rectangular sub-image of the region of interest (ROI). The CAD system analyzed properties of the ROI to distinguish malignant from benign tumors. The proposed system utilized intensity variation and textural information from the ROI sub-images as features to diagnose breast tumors. The US image database consists of 600 images of pathologically proven benign breast tumors from 370 patients, and carcinomas from 230 patients. The breast US images were acquired from four different US systems: Aloka SDD 1200, ATL HDI 3000, ATL HDI 5000 and GE LOQIC 700

* This work was supported by the National Science Council, Taiwan, under Grants NSC-92-2213-E-029-022.

scanners. Only one image extracted from each patient was used in the database. Figure 1(a) illustrates a real-time digitized monochrome US image. Figure 1(b) presents an exacted ROI of the tumor.



(a)

(b)

Fig. 1. (a) A 736 × 556 full breast US image, (b) the ROI rectangle is captured with a resolution of 169 × 135 pixels, approximately 2.91cm × 2.33 cm in size

2.1 Textural Analysis

Different US systems may produce inconsistent images. In order to obtain similar intensity contrast in the image sets, the breast US images require a preprocessing procedure. Histogram equalization is a mathematical process that could enhance the image contrast and reduce variation between images from different US systems. Figure 2 shows the ROI sub-image from the original US image, the equalized sub-image and the corresponding histogram.

The textural variation between benign and malignant in the US image is an efficient feature to classify breast tumors. The proposed CAD system utilized the correlation between neighboring pixels within the images as features to classify breast tumor. The auto-covariance coefficients between pixel (i, j) and pixel $(i+\Delta m, j+\Delta n)$ in an image with size $M \times N$ cab be defined as

$$\gamma(\Delta m, \Delta n) = \frac{A(\Delta m, \Delta n)}{A(0,0)} \qquad (1)$$

and

$$A(\Delta m, \Delta n) = \frac{1}{(M-\Delta m)(N-\Delta n)} \sum_{x=0}^{M-1-\Delta m} \sum_{y=0}^{N-1-\Delta n} \left| (f(x,y)-\bar{f})(f(x+\Delta m, y+\Delta n)-\bar{f}) \right|, \qquad (2)$$

where $\bar{f}$ is the mean value of $f(x, y)$. This paper utilized the coefficients for each breast tumor US image as the textural features vector. The PCA is used to reduce the dimension of the vector and then to distinguish the differences between benign and malignant tumors.

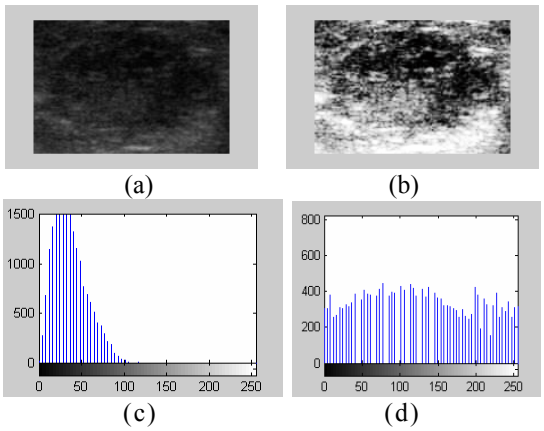


Fig. 2. (a) Original ROI, (b) histogram equalized ROI, (c) histogram of Original ROI and (d) histogram of preprocessed ROI

2.2 Principal Component Analysis

The idea behind PCA is to find another more applicable representation for the image vectors and reduces the dimension of the original representative vectors. The modified normalized auto-covariance matrix is used as the feature vector for representing each tumor ROI sub-image. The mathematical steps used to determine the principal components of a training set are given as below: Assume that there are N breast tumor images in the training set. The average feature vector m from the training set is given by

$$m = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i , \qquad (3)$$

where $\vec{x}_i$ is formed from normalized auto-covariance matrix. In this work, both Δm and Δn are 7, so processing a tumor image produces a 7 × 7 auto-covariance matrix (49-D auto-covariance coefficients). Because the value of $\gamma(0, 0)$ is always 1, except for the element $\gamma(0, 0)$, other coefficients are formed as a 48-D textural feature vector. An $N \times N$ matrix O is formed, whose elements O_{ij} are given by the inner product of feature vectors $(x_i - m)$ and $(x_j - m)$. Let ν_n and λ_n be the eigenvectors and the eigenvalues of O, respectively

$$O_{N\times N}=\begin{bmatrix}(x_1-m)\cdot(x_1-m) & \cdots & (x_1-m)\cdot(x_N-m)\\ \vdots & \ddots & \vdots\\ (x_N-m)\cdot(x_1-m) & \cdots & (x_N-m)\cdot(x_N-m)\end{bmatrix}_{N\times N}. \quad (4)$$

These eigenvectors determine linear combinations of the training set images to form the basis set of images u_i. The best characteristics of the variation in the training images can be represented by u_i:

$$u_i=\sum_{k=1}^{N} v_{ik}(\vec{x}_k-m), \quad (5)$$

for i = 1, 2, . . . N. The basis set images associated with the largest eigenvalue capture most of the information of the training set images. Then each training image x_k can be approximated with a linear combination of these basis set of u_i .The approximation equation is defined as

$$x_k \approx \sum_p \omega_p \mu_p . \quad (6)$$

The coefficients w_p are the new feature vectors representing the image x_k. A new query image, q_i, similarly can be approximated with the same linear combination and computed the coefficients w_q. The best way to find the most similar images from the training set to the new query image, q_i, is defined in terms of the Euclidean distance of the coefficients w_q and w_p. The match images will be selected from the training set depending on whose coefficients are the closest (in the Euclidean sense).

3. BREAST CANCER DIAGNOSIS

The diagram of the proposed CAD system was shown in Fig.3. The CAD system retrieves the first k US tumor images with smaller Euclidean distances from the training set. Depending on the *DS* value of those retrieved tumor US images, the new query image would be diagnosed as benign or malignant. The *DS* value is defined as

$$DS=\sum_{i=1}^{k}\frac{k-i+1}{Sum_k}\times Turmor\_class \quad (7)$$

and

$$Turmor\_class=\begin{cases}1, \text{if the retrieved result is malignant}\\ 0, \text{if the retrieved result is benign}\end{cases}, \quad (8)$$

where Sum_k is the summation of 1 to k . If the evaluated *DS* value is larger than a predefined threshold *Th*, the tumor is classified as malignant tumor. Conversely, if the evaluated *DS* value is less than a *Th*, the tumor is classified as benign tumor. Each retrieved image will be assigned a weight value. The values of weights are determined by the corresponding retrieved order.

4. EXPERIMENT RESULTS

The k-fold cross-validation method is used to estimate the performance of the proposed CAD system. In the simulations, k is 10 and each group has 60 US images. Random choose one of ten groups as test set, the rest of groups are as training set. Keep on recurring until every group is taken turns to be test set. Accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are used to evaluate the performance of diagnosis results. Table 1 shows the performance of retrieving different numbers of tumor US images from training set. With retrieving the first nine US images, the proposed system achieves a satisfied result. Table 2 illustrated the comparison with different threshold values when retrieves the US images from database. Table 3 demonstrates the diagnosis performance of retrieval nine US images.

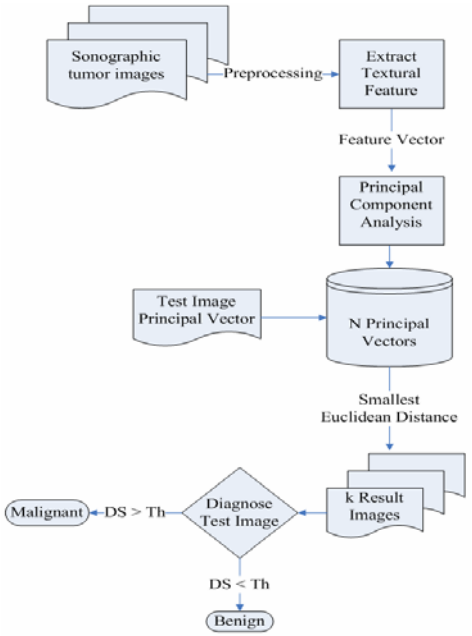


Fig. 3. Flow chart of the proposed CAD system

Table 1. The performance of retrieving number of k different US images (denotes RN-k).

	RN-5	RN-7	RN-9	RN-15
Accuracy	90.2%	91.3%	91.2%	89.7%
Sensitivity	96.1%	96.5%	97.0%	96.1%
Specificity	86.5%	88.1%	87.6%	85.7%
PPV	81.5%	83.5%	82.9%	80.7%
NPV	97.3%	97.6%	97.9%	97.2%

TP = true-positive; TN = true-negative; FP = false-positive; FN = false-negative

Accuracy = (TP+TN)/(TP+TN+FP+FN); sensitivity = TP/(TP+FN); specificity = TN/(TN+FP); PPV = TP/(TP+FP); NPV = TN/(TN+FN)

Table 2. The performance of RN-9 with different threshold values.

Threshold	TP	TN	Accuracy (%)	Sensitivity (%)
0.5	209	342	91.8%	90.9%
0.4	215	330	90.8%	93.5%
0.3	223	324	91.2%	97.0%
0.2	226	303	88.2%	98.3%

Table 3. Classification of breast nodules by proposed CAD system with *Th* = 0.3 and number of retrieval images is nine.

US image classification	Benign	Malignant
Benign(DS＜0.3)	TN 324	FN 7
Malignant(DS≧0.3)	FP 46	TP 223

5. CONCLUSION

In this paper, we proposed a practical CAD system for diagnosing breast cancer on sonogram. The proposed system achieved very good diagnosis performance. The US images in the database were acquired from the four different US systems. The simulations show that the proposed system performed well in the different US systems. The PCA was utilized to reduce dimensions of feature vector and avoided the perplexity training procedure. Diagnosis with the retrieval technique improved the efficiency of the CAD system and conserved diagnostic time. The proposed system provides a second opinion for physician to discriminate benign from malignant breast tumors. Historical cases can be directly added into the reference database without training again. With the growth of the database, the new cases will be collected and used as references while performing diagnoses.

6. REFERENCES

[1] "Breast Cancer Facts & Figures 1997-1998," *American Cancer Society*, 1999.

[2] A.T. Stavros, D. Thickman, C.L. Rapp, M.A. Dennis, S.H. Parker, and G.A. Sisney, "Solid Breast Nodules - Use of Sonography to Distinguish Benign and Malignant Lesions," *Radiology*, vol. 196, no. 1, pp. 123-134, July 1995.

[3] D.R. Chen, R.F. Chang, W.J. Kuo, M.C. Chen, and Y.L. Huang, "Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks," *Ultrasound Med. Biol.*, vol. 28, no. 10, pp. 1301-1310, Oct. 2002.

[4] D. Chen, R.F. Chang, and Y.L. Huang, "Breast cancer diagnosis using self-organizing map for sonography," *Ultrasound Med. Biol.*, vol. 26, no. 3, pp. 405-411, Mar. 2000.

[5] D.R. Chen, R.F. Chang, Y.L. Huang, Y.H. Chou, C.M. Tiu, and P.P. Tsai, "Texture analysis of breast tumors on sonograms," *Seminars in Ultrasound Ct and Mri*, vol. 21, no. 4, pp. 308-316, Aug. 2000.

[6] D.R. Chen, R.F. Chang, and Y.L. Huang, "Computer-aided diagnosis applied to US of solid breast nodules by using neural networks," *Radiology*, vol. 213, no. 2, pp. 407-412, Nov. 1999.

[7] B.S. Garra, B.H. Krasner, S.C. Horii, S. Ascher, S.K. Mun, and R.K. Zeman, "Improving the Distinction Between Benign and Malignant Breast-Lesions - the Value of Sonographic Texture Analysis," *Ultrasonic Imaging*, vol. 15, no. 4, pp. 267-285, Oct. 1993.

[8] S. Haykin, *Neural Networks: a comprehensive foundation*, 2 ed. NJ: Prentice Hall, 1999.

[9] R.C. Gonzalez and R.E. Woods, *Digital image processing*, 2 ed. Massachusetts: Addison Wesley, 2002.

[10] B. Maess, A.D. Friederici, M. Damian, A.S. Meyer, and W.J.M. Levelt, "Semantic category interference in overt picture naming: Sharpening current density localization by PCA," *Journal of Cognitive Neuroscience*, vol. 14, no. 3, pp. 455-462, Apr. 2002.

[11] U. Sinha and H. Kangarloo, "Principal component analysis for content-based image retrieval," *Radiographics*, vol. 22, no. 5, pp. 1271-1289, Sept. 2002.

[12] S. Costa and S. Fiori, "Image compression using principal component neural networks," *Image and Vision Computing*, vol. 19, no. 9-10, pp. 649-668, Aug. 2001.

[13] G.L. Gimelfarb and A.K. Jain, "On retrieving textured images from an image database," *Pattern Recognition*, vol. 29, no. 9, pp. 1461-1483, Sept. 1996.

[14] V.N. Gudivada and G.S. Jung, "An architecture for and query processing in distributed content-based image retrieval," *Real-Time Imaging*, vol. 2, no. 3, pp. 139-152, June 1996.

[15] V.N. Gudivada and V.V. Raghavan, "Content-Based Image Retrieval-Systems," *Computer*, vol. 28, no. 9, pp. 18-22, Sept. 1995.

[16] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837-842, Aug. 1996.

[17] W.J. Kuo, R.F. Chang, C.C. Lee, W.K. Moon, and D.R. Chen, "Retrieval technique for the diagnosis of solid breast tumors on sonogram," *Ultrasound in Medicine and Biology*, vol. 28, no. 7, pp. 903-909, July 2002.