Computer-Aided Diagnosis Applied to US of Solid Breast Nodules by Using Principal Component Analysis and Image Retrieval*

Yu-Len Huang†, Dar-Ren Chen‡ and Sheng-Hsiung Lin†
†*Department of Computer Science and Information Engineering Tunghai University, Taichung, Taiwan*
‡ *Department of General Surgery, Changhua Christian Hospital, Changhua, Taiwan*
ylhuang@mail.thu.edu.tw; dlchen88@ms13.hinet.net

Abstract*-This paper combines three useful textural features of ultrasound (US) images, i.e. block difference of inverse probabilities (BDIP), block variation of local correlation coefficients (BVLC) and auto-covariance matrix, to classify benign and malignant breast tumors. 1020 sonograms of region of interest (ROI) from 255 patients were used as case samples. Two-view sonogram (longitudinal and transverse view) and four different rectangular regions are utilized for each tumor analysis. The textural features always perform as a high dimensional vector. High dimensional vector is unfavorable to differentiate breast tumors in practice. 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. The proposed computer-aided diagnosis (CAD) system differentiates solid breast nodules with a relatively high accuracy in the US system and helps inexperienced operators avoid misdiagnosis.*

Keywords: ultrasound, principal component analysis, image retrieval, computer-aided diagnosis, textural analysis, breast cancer.

1. Introduction

Early detection and treatment of breast cancer is a useful way to increase the cure rate [1]. Accurate and reliable diagnostic procedure is important in the early diagnosis. Early treatment in time can prevent cancer cells spreading. Diagnosis of breast tumors frequently adopts mammography and sonography in clinical practice. Those modalities are supplied for physician to differentiate benign breast tumors from malignant lesions. The breast sonography and mammography can help physician to evaluate a breast mass in daily clinical practice. Breast sonography generally played a role as an auxiliary to mammography. However, the ultrasound (US) examination is more convenient and safer than mammography for patient in regularly physical examination. In 1995, Stavros et al. indicated the US technique is helpful to predict breast cancer more accurate [2]. The authors point out the US technique required an accomplished radiologist with extensive real-time evaluation, most clinical environment may not agree with the application. Physicians use mammography and sonography to diagnose breast cancer via visual experiences. 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 accuracy of diagnosis, the computer-aided diagnosis (CAD) system can be a second beneficial support for physician to make correct diagnosis. Because of a great deal of equivocal lesions need to be distinguished by biopsy per year, but biopsy is still a costly and invasive method. If the proposed CAD can increase the rate of positive findings for breast cancer, unnecessary biopsy may be avoidable, alleviate anxiety and control costs.

The textural variation in the US image has been found as a useful feature to identify benign and malignant tumours [3]. Chen et al. utilized the textural features in breast US images to differentiate between benign and malignant tumors by using the neural network (NN) classifier [4-7]. With the growth of the database, more and more information may be collected and used as reference cases while performing diagnosis. The NN-based diagnosis system must be retraining for adding historical cases into the database. In order to solve this problem, this paper proposed an image retrieval diagnosis system that distinguished malignant from benign masses on the textural similarity of the breast US image. The proposed CAD system utilized effortless textural features, i.e. block difference of inverse probabilities (BDIP), block variation of local correlation coefficients (BVLC) and auto-covariance matrix, to identify breast tumor. The textual feature vector is always in a high dimensional space. Performing the feature vector directly is unfavourable to identify breast tumors by image retrieval. Thus we perform the principal component analysis (PCA) [8-9] to diminish the dimension of the feature vector. The

* This work was supported by the National Science Council, Taiwan, under Grants NSC93-2213-E-029-014.

original textural feature vector will be mapping into principal vector with a lower dimension. The transformed vector, the principal vector, is used as new textural feature to retrieve images from database based on similarity measure of Euclidean distance. The retrieved images are supplied as the reference resources to identify benign and malignant lesions in the US image. The proposed CAD system achieves a good diagnostic performance by using image retrieval techniques with PCA on textural features.

2. Data Acquisition

1020 sonograms of region of interest (ROI) from 255 patients including 36 cancers, 57 fibrocystic nodules, 120 fibroadenomas and 33 cysts were used as case samples. The ultrasonic appearances were then correlated either with the fine needle aspiration, core biopsy or surgical findings. ROI was manual extracted by physicians. The ultrasonic images were captured at the transverse and longitudinal views for each tumor. The images were collected from June 1, 1998 to Apirl 31, 1999; the patients' ages ranged from 18 to 64 years; tumors were from 0.8 to 3.6 cm in size. Sonography was performed using an ALOKA SSD 1200 (Tokyo, Japan) scanner and a 7.5 MHz lineal transducer with freeze-frame capability. No acoustic standoff pad was used in any of the cases.

When a sonogram is performed, an analog video signal is transmitted from the VCR output of the scanner to a portable notebook computer; the data is then digitized by a frame grabber Video CATcher (from the Top Solution Technology Co.) that is connected to the printer port of the computer. The capturing resolutions of the portable computer and the external frame grabber are 736 × 566 pixels for an NTSC video screen picture. The monochrome ultrasonic image is quantized into 256 gray levels. Figure 1 presents a tumor in the different views.

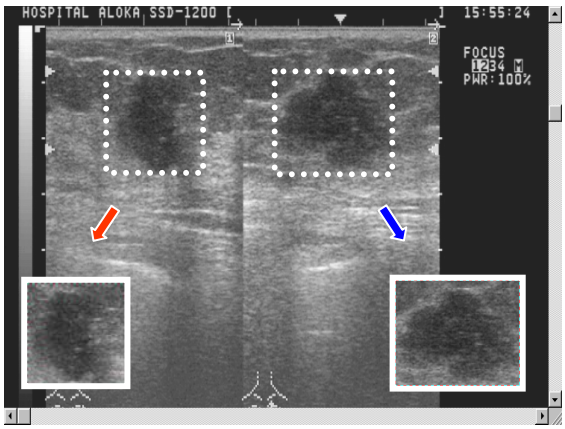

Figure 1. A breast tumor located in the different views. (with a resolution of 58 × 58 pixels in a 1cm × 1cm rectangle)

Four different rectangular regions from the two sonograms are used for the analysis of each tumor. An example is shown in Fig. 2. There were: (a) two regions that extended beyond the lesion margins in all directions by 1-2 mm for both transverse and longitudinal views of a tumor; (b) the largest rectangular region that would fit inside the lesion for both transverse and longitudinal views of a tumor.

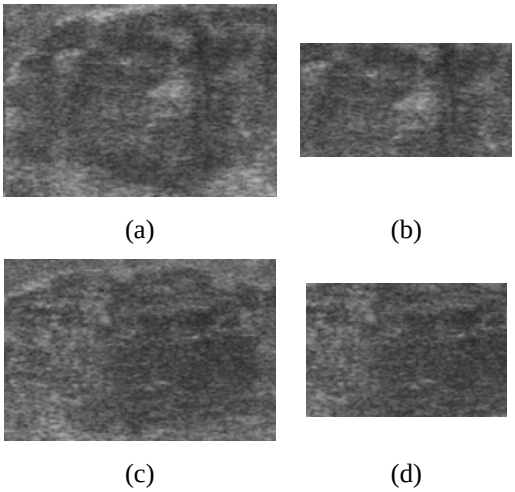

Figure 2. An example of the four ROI subimages for a tumor: (a) full longitudinal ROI (LA) subimage, (b) inside longitudinal ROI (LB) subimage, (c) full transverse ROI (TA) subimage, and (d) inside transverse ROI (TB) subimage.

3. Image Analysis

An ultrasonic image consists of many points with different values of gray level intensity. Different tissues have significantly different textures. The block difference of inverse probabilities (BDIP) and block variation of local correlation coefficients (BVLC) image features are proposed by Young et al. for content-based image retrieval [10]. The proposed CAD system adopts these two textural features and the autocorrelation matrix to differentiate benign breast tumors from malignant lesions. BDIP and BVLC are defined as following equations.

$$BDIP = M^2 - \frac{\sum_{(i,j)\in B} I(i,j)}{\max_{(i,j)\in B} I(i,j)},$$

where $I(i,j)$ is the intensity of pixel (i, j) and B is an $M \times M$ block. The larger value of BDIP would be if the larger variance of intensities in a block. In this paper, M is chosen to be 2.

$$\rho(k,l) = \frac{\frac{1}{M^2}\sum_{(i,j)\in B} I(i,j)I(i+k,j+l)\mu_{0,0},\mu_{k,l}}{\sigma_{0,0},\sigma_{k,l}},$$

$$BVLC = \max_{(k,l)\in O_4}[\rho(k,l)] - \min_{(k,l)\in O_4}[\rho(k,l)],$$

$$O_4 = \{(0,1),(1,0),(1,1),(1,-1)\}.$$

where $\mu_{0,0}$, $\sigma_{0,0}$ are the local mean and standard deviation of the block with size $M \times M$. The (k, l) term denotes four shift directions, they are -90°, 0°, -45°, 45° respectively. The $\mu_{k,l}$, $\sigma_{k,l}$ are the mean and standard deviation of the shifted block. The larger BVLC value means that the ingredients in the block are rough.

The 2-D normalized auto-correlation coefficients used to reflect the inter-pixel correlation within an image. The coefficients are further modified into a mean-removed version to generate the similar auto-covariance features for images with different brightness but with a similar texture. The modified auto-covariance coefficients between pixel (i, j) and pixel (i+Δm, j+Δn) in an image with size $M \times N$ cab be defined as

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

and

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

where $\bar{f}$ is the mean value of $f(x,y)$. The dimension of the auto-covariance matrix can be any size of images. In this study, Δm and Δn are both 7, after the processing of texture analysis, a 7 × 7 auto-covariance matrix is obtained for each image. Because the value of $\gamma(0,0)$ is always zero. The first element in matrix of every US image will be discarded.

4. Principal Component Analysis

PCA is a well-known statistical processing technique that can reduce redundancy by projecting the original data over a proper basis. The result which PCA provided is a more applicable and diminished dimension vector. The following is the mathematical procedure of determine the principal components of a training set. We can view the previous textural features of a ROI subimage as a vector. Because the first element of the feature vector is always 1, the first element will be discarded and the rest elements of four ROI subimages for each case can be combined as a 192 dimension feature vector. Suppose that there are N feature vectors in the training set. The average feature vector m of the training set is given by

$$m = \frac{1}{N}\sum_{i=1}^{N} \vec{x}_i$$

where $\vec{x}_i$ is the high dimension feature vector of the ith ROI subimage in the training set. The linear combinations of the eigenvectors the training set form the basis set of vectors u_i. We can represent the best characteristics of the variation in the training vectors as principle component u_i:

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

for i = 1, 2, …, N. The basis set vectors formed from largest eigenvalues contain most of the information of the feature vectors in the training set. The percent of the total variability explained by each u_i can be figured out. Generally, we use first p principal components which exceed 90% of the total variance of the original vectors to approximately project the original feature vector x_k into a new p-dimensional feature vector. The approximation equation is defined as

$$x_k \approx \sum_p \omega_p \mu_p,$$

where w_p are the new feature vectors representing the x_k. The textural feature vector of a queried ROI subimage, q_i, can be approximated by the same linear combination and coefficients w_q. The coefficients w_p are the new feature vectors representing the x_k. The textural feature vector from a query ROI subimage, q_i, can be approximated with the same linear combination and computed the coefficients w_q. An analysis was performed to assess the effects of the new feature vector for the US image database. In this study, we found that the ideal p value is 10, so each original 192-D textural feature vector was reduced by PCA into a 10-D new feature vector.

5. Image Retrieval for Breast Cancer Diagnosis

Firstly, the distance of the coefficients w_q and w_p need to be computed. Similar images were selected from database depending on the criterion of Euclidean distance. The proposed CAD system retrieves the first L tumor images with the smallest Euclidean distance. The queried image would be diagnosed as benign or malignant depending on the DS value of those retrieved images. The DS value is defined as

$$DS = \sum_{i=1}^{L} Weight_i \times Tumor\_class_i$$

$$Weight_i = \frac{L - i + 1}{\sum_{j=1}^{L} j},$$

$$Tumor\_class_i = \begin{cases} 1, \text{if the retrieved image } i \text{ is malignant case} \\ 0, \text{if the retrieved image } i \text{ is benign case} \end{cases}$$

The weight of each retrieved images is determined by the retrieved order. A cut-off threshold *Th* was predefined to separate benign and malignant tumor. If *DS* value is greater than *Th*, the tumor is classified as malignant one. Otherwise, the tumor is benign.

6. Results

This study classifies the benign and malignant tumors based on retrieved US images. The *k*-fold cross-validation method [11] is used to estimate the performance of a CAD system. The proposed CAD system was trained and tested with *k*-fold cross validation (*k* = 10) methods to recognize the malignant or benign tumors. The 255 cases (1020 US images) in the database randomly divided into *k* groups. The performance of the proposed CAD system was also analyzed with ROC curve. With a cut-off threshold of 0.15 and the number of retrieval image is 7, the proposed CAD system correctly identifies 31 of 36 malignant tumors and 188 of 219 benign tumors. The proposed CAD system achieved an area under the ROC curve of 0.9253±0.007, as shown in Fig. 3. Table 1 lists the number of misdiagnosed cases at threshold = 0.15 for each test set. The accuracy of proposed CAD system for malignancy, the sensitivity, the specificity are illustrated in Table 2.

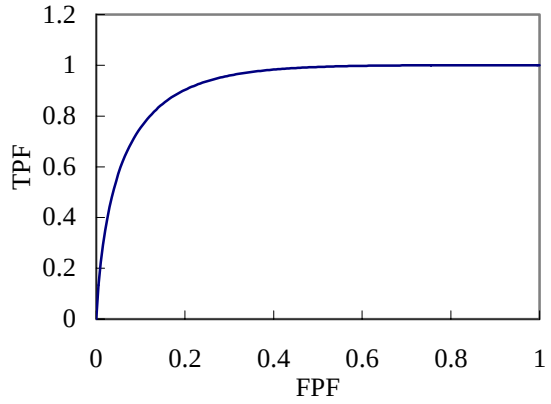

Figure 3. The diagram of the ROC curve for the retrieval technique is employed in classifying of malignant and benign tumors (the A_z value for the ROC curve is 0.9253 ± 0.007).

Table 1. The number of misdiagnosed cases of the proposed CAD system at threshold = 0.15 and the number of retrieved image is 7 for each test set.

Test Set	Malignant Cases	Benign Cases
1	1/4	1/21
2	0/4	2/22
3	0/4	2/22
4	0/4	7/22
5	2/4	1/22
6	0/4	7/22
7	0/3	2/23
8	1/3	3/22
9	1/3	3/22
10	0/3	4/22

7. Discussion

Texture features are helpful to classify masses and normal tissue on sonography, the image retrieval technique provides a potentially useful tool for the sonographic decision support. This study proposes an efficient and effective CAD system with multi-view sonograms to distinguish between benign and malignant tumors. The proposed CAD system diagnoses breast tumors using inter-pixel correlations within the four ROI subimages. Based on the experimental results, the proposed CAD system performs differential diagnosis very well. These results confirm that benign and malignant tumors can be classified using texture features in multi-view digital US images. From the satisfactory specificity and sensitivity of results, the proposed system is expected to be a useful computer-aided diagnostic tool for differentiating between benign and malignant cases using sonograms, and could avoid misdiagnosis and reduce the number of unnecessary surgical biopsies.

Table 2. The performance of the proposed CAD system at threshold = 0.15 and the number of retrieval image is 7.

	Benign	Malignant
DS value < 0.15	TN 188	FN 5
DS value >= 0.15	FP 31	TP 31
Total	219	36

Note: TN: True Negative, FN: False Negative, FP: False Positive, TP: True Positive
(1) Accuracy = (TP+TN)/(TP+TN+FP+FN) = 85.9%
(2) Sensitivity = TP/(TP+FN) = 86.1%
(3) Specificity = TN/(TN+FP) = 85.8%

References

[1] "Breast Cancer Facts & Figures 2001-2002," *American Cancer Society*, 2003.

[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] 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.

[4] 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.

[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, "Breast cancer diagnosis using self-organizing map for sonography," *Ultrasound Med. Biol.*, vol. 26, no. 3, pp. 405-411, Mar. 2000.

[7] 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.

[8] I. T. Jolliffe, Principal Component Analysis. New York: Springer-Verlag, 1986.

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

[10] D. C. Young and Y. S. Sang, “Image retrieval using BDIP and BVLC moments,” *IEEE Trans. on circuits and systems for video technology*, vol. 13, no. 9, pp. 951-957, Sept. 2003.

[11] S.M. Weiss and I. Kapouleas, "An empirical comparison of pattern recognition neural nets and machine learning classification methods," *Proc 11th Int Joint Conf Artificial Intelligence*, pp. 234-237, 1989.