

Image retrieval with principal component analysis for breast cancer diagnosis on various ultrasonic systems

Y.-L. HUANG*, S.-J. KUO†, C.-S. CHANG‡, Y.-K. LIU*, W. K. MOON§ and D.-R. CHEN†

*Department of Computer Science and Information Engineering, Tunghai University, Taichung, Departments of †General Surgery and ‡Medical Oncology, Changhua Christian Hospital, Changhua, Taiwan and §Department of Diagnostic Radiology, College of Medicine, Seoul National University Hospital, Seoul, South Korea

KEYWORDS: breast cancer; computer-aided diagnosis; image retrieval; principal component analysis; textural analysis; ultrasound

ABSTRACT

Objectives We present a computer-aided diagnostic (CAD) system with textural features and image retrieval strategies for classifying benign and malignant breast tumors on various ultrasonic systems. Effective applications of CAD have used different types of texture analysis. Nevertheless, most approaches performed in a specific ultrasonic machine do not indicate whether the technique functions satisfactorily for other ultrasonic systems. This study evaluated a series of pathologically proven breast tumors using various ultrasonic systems.

Methods Altogether, 600 ultrasound images of solid breast nodules comprising 230 malignant and 370 benign tumors were investigated. All ultrasound images were acquired from four diverse ultrasonic systems. The suspicious tumor area in the ultrasound image was manually chosen as the region-of-interest (ROI) subimage. Textural features extracted from the ROI subimage are supported in classifying the breast tumor as benign or malignant. However, the textural feature always behaves as a high-dimensional vector. In practice, high-dimensional vectors are unsatisfactory at differentiating breast tumors. This study applied the principal component analysis (PCA) to project the original textural features into a lower dimensional principal vector that summarized the original textural information. The image retrieval techniques were employed to differentiate breast tumors, according to the similarities of the principal vectors. The query ROI subimages were identified as malignant or benign tumors according to characteristics of retrieved images from the ultrasound image database.

Results Using the proposed CAD system, historical cases could be directly added into the database

without a retraining program. The area under the receiver–operating characteristics curve for the system was 0.970 ± 0.006 .

Conclusion The CAD system identified solid breast nodules with comparatively high accuracy in the different ultrasound systems investigated. Copyright © 2005 ISUOG. Published by John Wiley & Sons, Ltd.

INTRODUCTION

Our group has applied textural features in breast ultrasound images to differentiate between benign and malignant tumors with neural network (NN) classifiers^{1–3}. Textural variation in the ultrasound image has been deemed a useful characteristic for distinguishing benign and malignant tumors^{4,5}. However, the NN training process is prolonged and diagnostic performance normally relies on the initial parameter setting⁶. A common weakness of computer-aided diagnostic (CAD) systems employing texture analysis is that they only work effectively in a specific ultrasonic system. However, with the rapid development of ultrasound technologies, numerous different ultrasonic systems are now employed in medical diagnosis. With the growth of the database, more information may be gathered and employed as reference cases for diagnostic support. The NN-based diagnosis system had to retrain for loading historical cases into the database. To resolve this difficulty, Kuo *et al.*⁷ designed an image-retrieval diagnosis system utilizing the co-occurrence matrix determined by the ultrasound images to distinguish benign and malignant tumors. Although these CAD systems offered satisfactory diagnostic performance, the weighting coefficients of the

Correspondence to: Dr D.-R. Chen, Department of General Surgery, Changhua Christian Hospital, 135 Nanhsiao Street, Chunghua, Taiwan 500 (e-mail: dlchen88@ms13.hinet.net)

Accepted: 4 March 2005

feature parameters still necessitated extensive evaluations. The parameters of the CAD system need adjusting for various ultrasonic systems with dissimilar resolutions and mechanism settings.

Recently, the quantity of digital images has expanded vastly for image database applications such as digital libraries, picture archiving and communications systems, geographic information systems and numerous others. With enlarged databases, content-based image queries and retrievals are crucial for finding the desired images automatically from the image database^{8,9}. Effective content-based image retrieval approaches must aim at characteristics of different types of images^{10,11}. Various content-based image retrieval systems, such as IBM's QBIC project, provide image retrieval capacity to automatically index and query images. This study employed image retrieval strategies to distinguish malignant from benign masses on the textural resemblance of the breast ultrasound image. The proposed CAD system utilized a facile textural feature (i.e. auto-covariance matrix) to identify breast tumor. However, the textural feature vector is consistently in a high-dimensional space. Performing the feature vector directly is unsatisfactory for identifying breast tumors by image retrieval. Thus, this study performed a dimension reduction method, principal component analysis (PCA)^{12,13}, to substitute a large set of feature vectors for a smaller set of new vectors. The original textural feature vector was mapped into principal vector with a lower dimension¹⁴. The projected vector, the principal vector, summarized the original vector with fewer dimensions and employed new textural features to retrieve images based on similarity measure of Euclidean distance (the shortest straight-line distance between two vectors). The retrieved images were supplied as reference resources for identifying benign and malignant lesions in the ultrasound image.

METHODS

Normally, a physician can readily pinpoint a tumor in a sonographic image by the tumor shape and the contrast of internal echoes. Automatic tumor segmentation on an ultrasound image is difficult. No satisfactory approaches appear to exist to date, to the authors' knowledge. Thus the physician manually extracted the rectangular subimage of the region-of-interest (ROI) in this study. The rectangular ROI included around 0–5 mm extension from the tumor border. The proposed system employed intensity variation and textural information from the ROI subimages as features with which to diagnose breast tumors.

Data acquisition

The ultrasound image database comprised 600 images of pathologically proven benign breast tumors from 370 patients and carcinomas from 230 patients (tumor size > 0.8 cm in all cases). The ultrasound images were captured at the largest diameter of the tumor. The breast ultrasound image databases contain only histologically

confirmed cases (either by fine-needle aspiration, core-needle biopsy or open biopsy) recorded from 1 August 1997 to 31 May 2000. In our practice, fine-needle aspiration was done for C3 cases and core-needle biopsy (in most cases)/excision biopsy for C4 or C5 cases. The breast ultrasound images were acquired from the following various ultrasonic systems:

1. SDD 1200 (Aloka, Tokyo, Japan): 226 digitized tumor images (69 malignant and 157 benign).
2. HDI 3000 (ATL, Bothell, WA, USA): 256 digital tumor images (125 malignant and 131 benign).
3. HDI 5000 (ATL): 55 digital tumor images (18 malignant and 37 benign).
4. LOGIQ 700 (GE, Waukesha, WI, USA): 63 digital tumor images (18 malignant and 45 benign).

The analog signals obtained by the SDD 1200 system from the VCR output of the scanner were transmitted to a frame grabber, Video CATcher (Top Solution Technology, Taipei, Taiwan). Then every monochrome ultrasound image was quantized into eight bits with 256 gray levels. The digital images were gathered before biopsy using the HDI 3000 system with an L10-5 small-part transducer, which is a linear-array transducer with a frequency of 5–10-MHz and a scan width of 38 mm. During the ultrasound scanning, no acoustic standoff pad was used. Patients were used for only one database image each. All images were supplied by the coauthors (W.K.M and D.R.C.) while the ROIs were chosen by one of the authors (D.R.C.). Throughout this study, only the ROI subimages were employed to evaluate the texture characteristics of benign and malignant lesions. The physician utilized the software package ProImage (Prolab, Taipei, Taiwan) to choose the rectangular ROI subimage manually, saving them in files for textural analysis. Figure 1a demonstrates a 640 × 480 real-time digitized monochrome ultrasound image. Figure 1b presents an exacted ROI with a resolution of 244 × 135 pixels, approximately 2.60 × 1.44 cm in size.

Textural analysis

Images acquired by various ultrasonic systems may result in influence of distinct gray level contrast and spatial resolution. For example, Figure 2 shows the ROI subimages of breast tumor acquired by various ultrasonic systems. Evidently, images from the various ultrasonic systems exhibit variations in contrast and resolution. To obtain similar contrast and resolution for the images in the ultrasound image database, a preprocessing adjustment would be performed on ultrasound images before analyzing the textural feature for the ROI subimages. For contrast adjustment, histogram manipulation can effectively enhance images. Histogram equalization¹⁵ is a mathematical procedure that could enhance the image contrast, reducing differences among images from various ultrasonic systems. Thus histogram equalization was implemented to preprocess all images in the

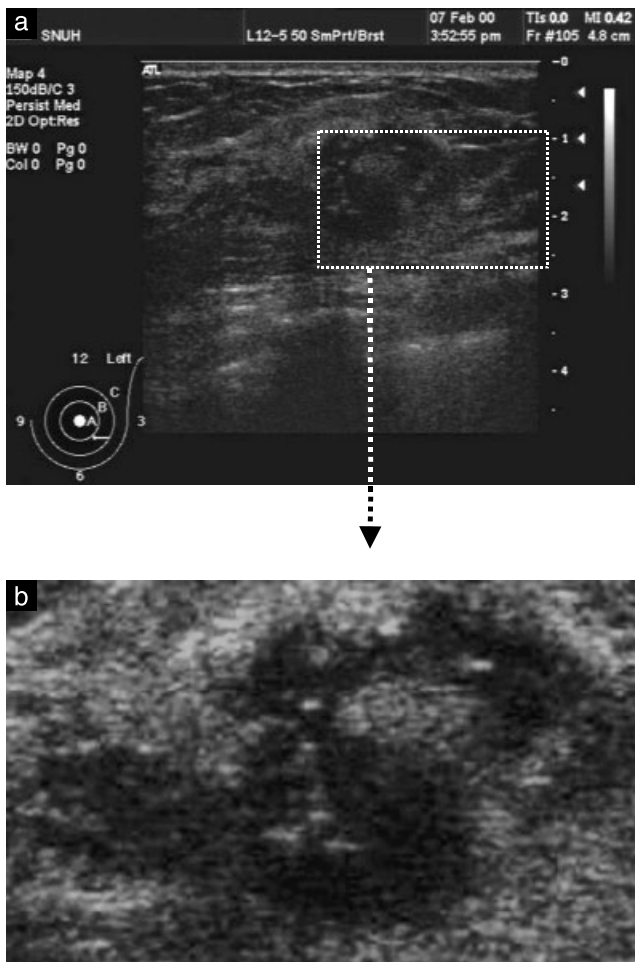


Figure 1 (a) A 640×480 full breast ultrasound image and (b) the region-of-interest subimage captured with a resolution of 244×135 pixels, approximately 2.60×1.44 cm in size.

ultrasound image database. The histogram equalization is satisfactory because the technique automatically enhances digital images and the results from this technique are predictable. Figure 3 illustrates the ROI subimage of the original ultrasound image, the equalized image and the corresponding histogram. The resolution adjustment between ultrasound images from various ultrasonic systems applied the bi-cubic interpolation approach to adapt spatial resolution of ROI subimages. The contrast and resolution adjustment should be carried out prior to evaluating texture features.

The textural variation between benign and malignant in the ultrasound image is an effective feature for classifying breast tumors. The proposed CAD system exploits the correlation between adjacent pixels within images as features to classify breast tumor. The two-dimensional normalized auto-correlation coefficients¹⁶ were utilized to reflect the inter-pixel correlation within an image. The coefficients are further modified into a mean-removed version to create the comparable auto-covariance features for images with disparate brightness but with a similar texture. These auto-covariance coefficients have been found to be effective textural features in breast ultrasound images for differentiating between benign and malignant

tumors¹⁻⁴. The modified auto-covariance coefficients between pixel (i, j) and pixel $(i + \Delta m, j + \Delta n)$ in an image with size $M \times N$ can be defined as:

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

and

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

where \bar{f} is the mean value of $f(x, y)$. The size of the auto-covariance matrix was $\Delta m \times \Delta n$. These auto-covariance coefficients represent feature vector for each tumor ROI subimage. However, the dimension of the textural feature vector is proportional to the size of the auto-covariance coefficients matrix (dimension of $\Delta m \times \Delta n$). The textural feature vector is consistently in a high-dimensional space for catching the textural variety in images. Performing the high-dimensional vector directly is unsatisfactory when identifying breast tumors. The PCA was applied in this study to diminish the dimension of the feature vector, projecting the original feature vector into a lower dimensional principal vector. The principal vector was then deemed to be the new textural feature. The evaluation of PCA is described below.

PCA for vector dimension reduction

PCA is a conventionally adopted statistical analytical method that facilitates diminishing redundancy by projecting the original data over a proper basis. The idea behind PCA is to create a more pertinent representation for reducing the dimension of the original vectors. The mathematical steps to establish the principal components of a training set are detailed in the Appendix.

An analysis was performed on the effects of the new feature vector for the ultrasound database. Figure 4 confirms the first three principal components ($p = 3$) explain over 95% of the total variability in the standardized ratings. According to our results, the ideal p value is 3, so each original 48-dimensional textural feature vector was condensed by PCA into a new three-dimensional feature vector.

Breast cancer diagnosis by image retrieval

The common means of selecting the most similar images from the image database to the new query image is expressed by using the Euclidean distance of the coefficients w_q and w_p . The retrieved images are selected from the image database according to the Euclidean distance criterion. The proposed CAD system selected the first L tumor ultrasound images with the smallest Euclidean distances from the ultrasound image database. Consequent on the difference score (DS) value of those

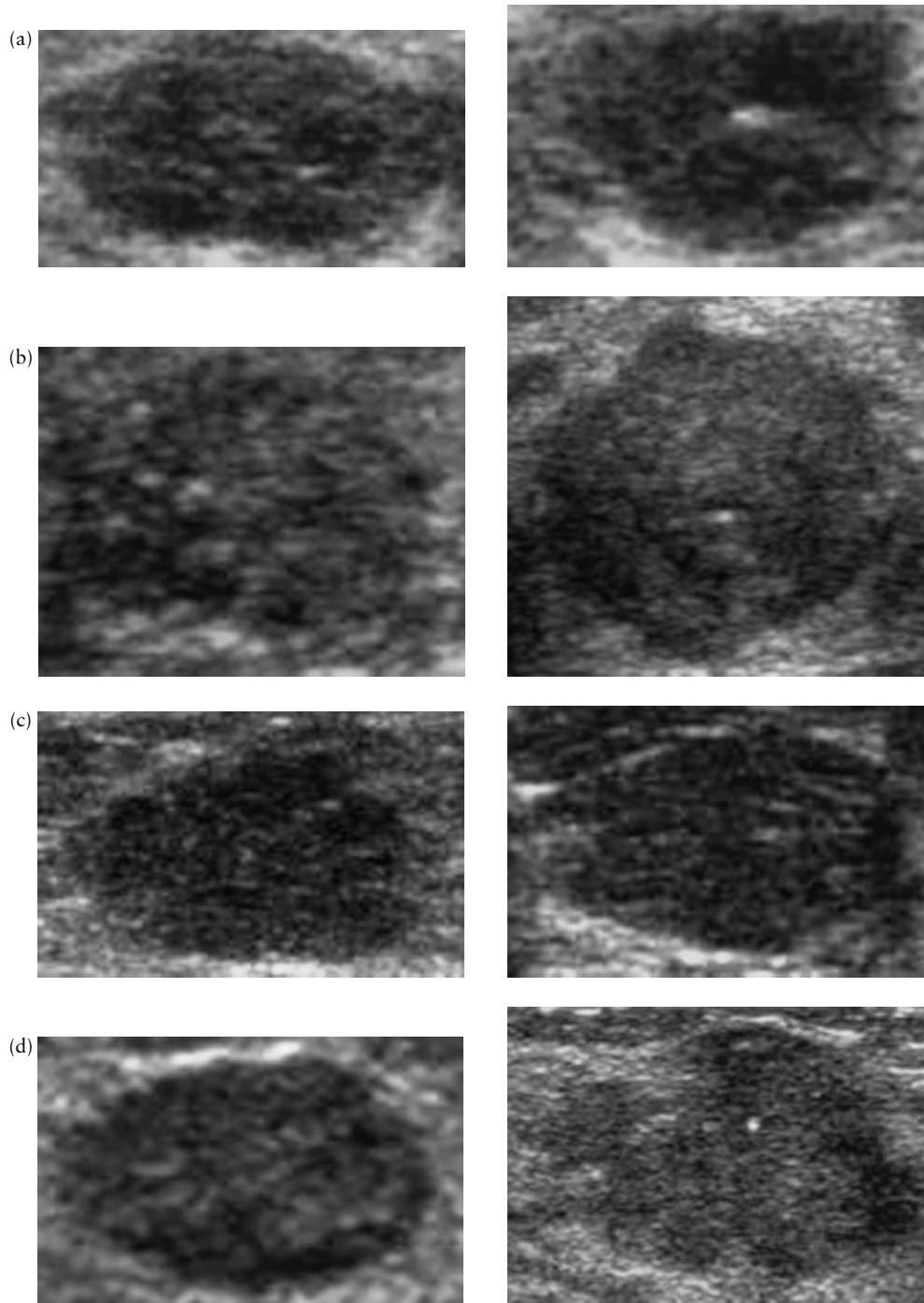


Figure 2 Region-of-interest subimages of breast lesions acquired by the following ultrasound systems: (a) SDD 1200, (b) HDI 3000, (c) HDI 5000 and (d) LOGIQ 700. The benign tumors are shown in the images on the left, and the malignant tumors in the images on the right.

retrieved images, the new query image would be diagnosed as a benign or malignant lesion. The *DS* value is defined as:

$$DS = \sum_{i=1}^L Weight_i \times Tumor_class_i, \tag{3}$$

$$Weight_i = \frac{L - i + 1}{\sum_{j=1}^L j}, \tag{4}$$

$$Tumor_class_i = \begin{cases} 1, & \text{if the retrieved image } i \text{ is malignant case} \\ 0, & \text{if the retrieved image } i \text{ is benign case} \end{cases} \tag{5}$$

Each retrieved image was assigned a weight value established by the corresponding selected order. A cut-off threshold *Th* was predefined as a demarcation line separating breast tumors. If the evaluated *DS* value was greater than *Th*, the tumor was diagnosed as malignant. Conversely, if the evaluated *DS* value was below *Th*,

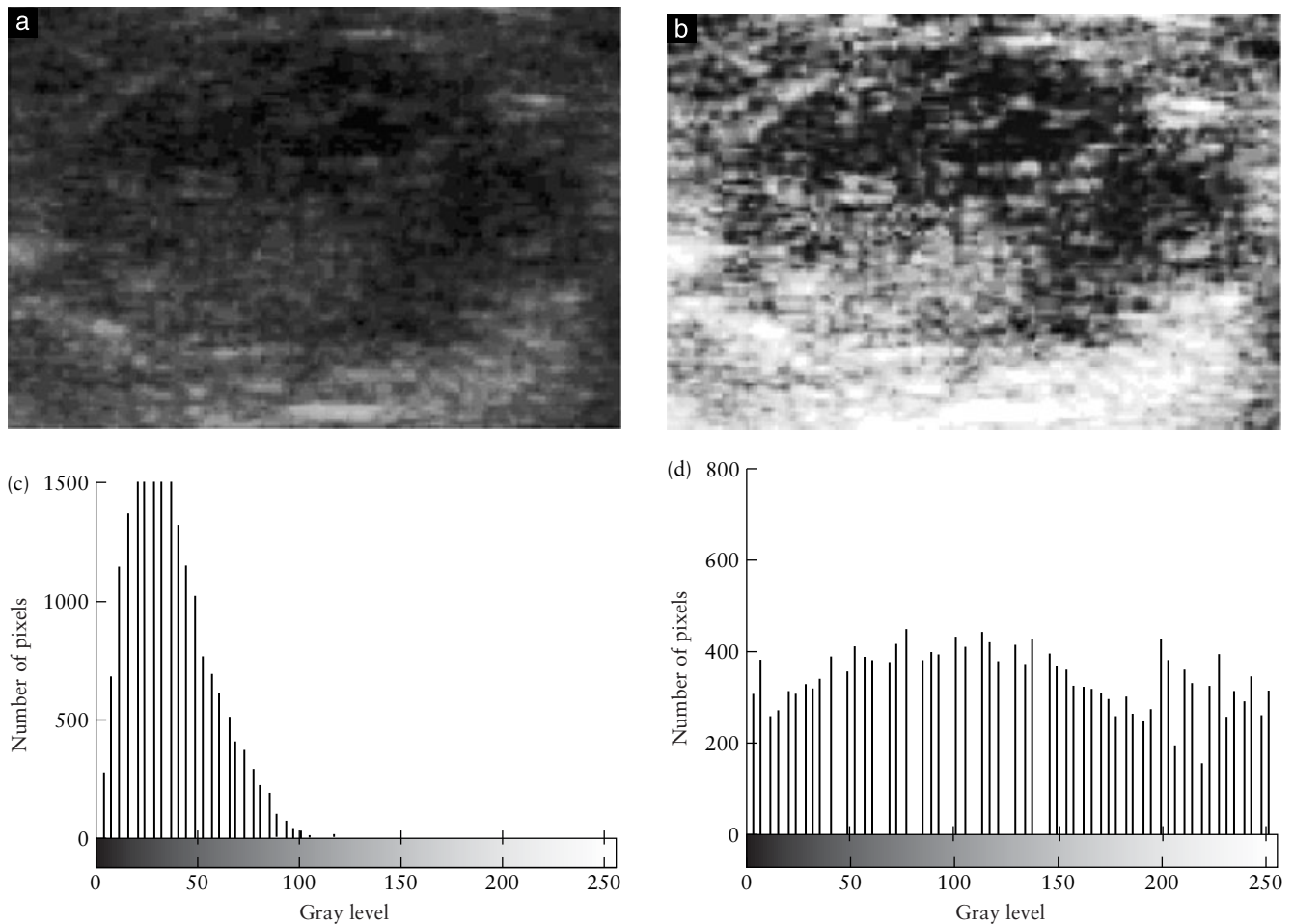


Figure 3 (a) An original region-of-interest (ROI) subimage, (b) the preprocessed image, (c) a histogram of the original ROI subimage and (d) a histogram of the preprocessed ROI.

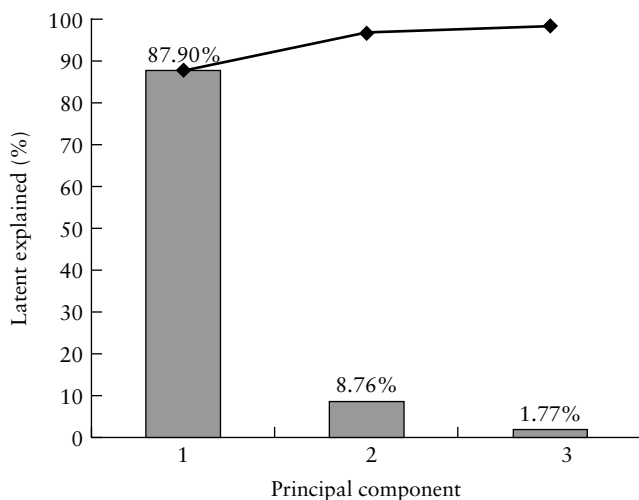


Figure 4 A graph of the percent variability explained by each principal component. The first three principal components from the textural feature vector explain over 98% of the total variability in the standardized ratings. The line at the top of the graph shows the cumulative sum of latent explained.

the tumor was diagnosed as benign. The flow chart of the proposed diagnostic approach is displayed in Figure 5.

Diagnosis evaluations

The k -fold cross-validation method¹⁷ was used to evaluate the performance of the proposed CAD system. The 600 ultrasound images in the database randomly divided into k groups. The first group was excluded and the other $(k - 1)$ groups functioned as the training set. The second group acted as a testing group while the ultrasound images in the remaining nine groups were trained. This process was repeated until each k group in turn became a testing group. Two performance measures were applied to gauge the performance of the diagnostic system. One measure encompassed diagnostic accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). The other measure was the A_Z value, which was calculated by the receiver–operating characteristics (ROC) curves (software package LABROC1 by Professor C. E. Metz, University of Chicago, Chicago, IL, USA). The area A_Z under the ROC curve is an index of the quantitative measure of the overall performance of a diagnostic system. The A_Z value could therefore compare performance using different methods to clearly distinguish positive and negative findings of breast tumors. The simulations were made on a single CPU Intel Pentium-4® 2.4

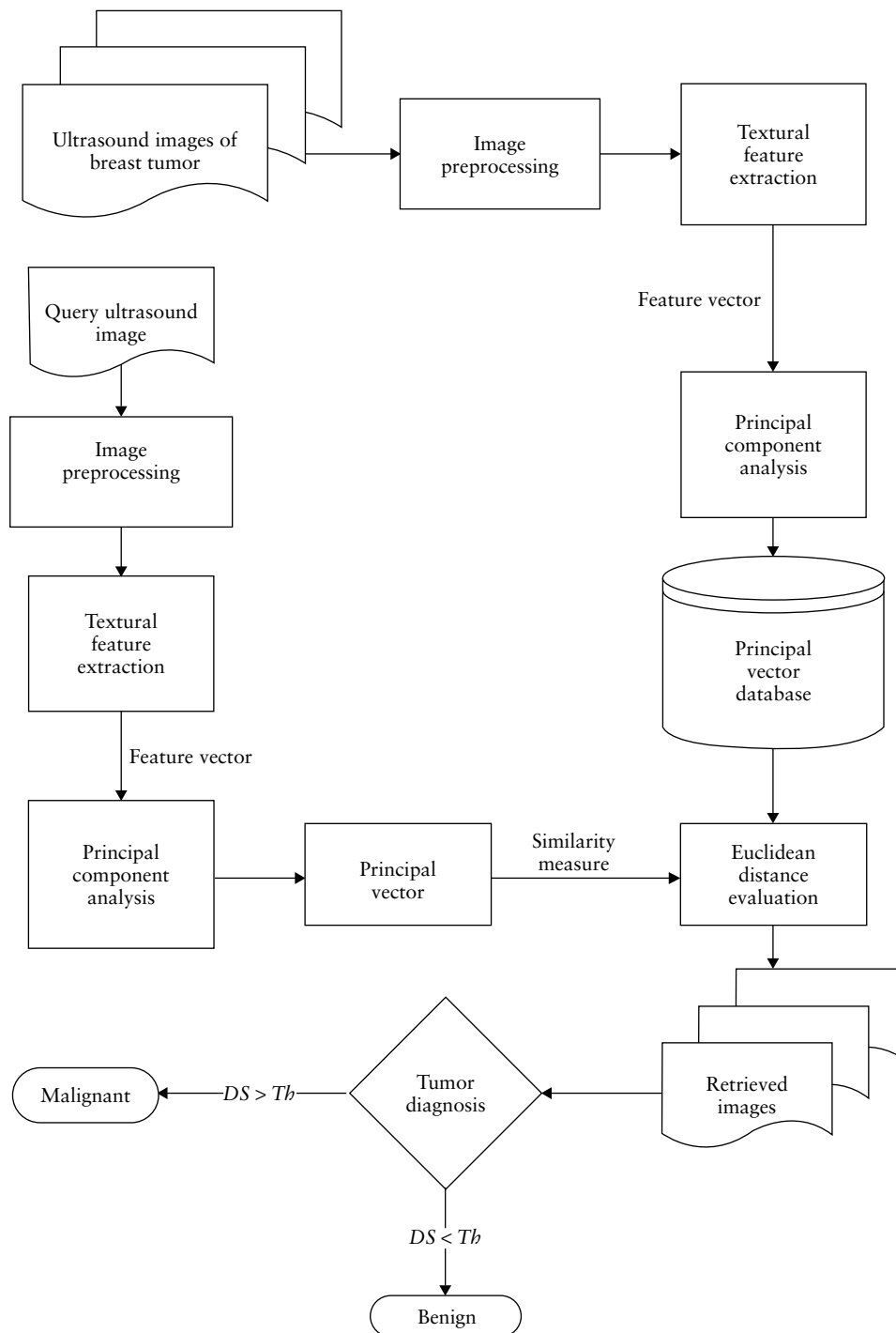


Figure 5 Flow chart of the proposed computer-aided diagnostic system.

GHz personal computer with Microsoft Windows XP® operating system.

RESULTS

The k was chosen as 10 in the simulations and each group included 60 ultrasound images. Figure 6 shows the A_Z value and diagnostic accuracy gained with the proposed CAD system ($n = 5, 7, 9, 11$ and 13). The retrieval performance rates of different numbers of tumor ultrasound images were comparable. Table 1 compares

the retrieval of different numbers of tumor ultrasound images for differentiating benign and malignant tumors. Although the results of all these five sets with different n values were satisfactory, the RN_9 set demonstrated good performance on average. Figure 7 shows the ROC curve for the proposed CAD system. The proposed system with RN_9 achieves $A_Z = 0.970 \pm 0.006$. Table 2 compares different threshold cut-off values for RN_9 . The ideal cut-off value Th of 0.3 was the best choice. Table 3 summarizes the diagnostic performance for RN_9 . To verify the practicality of the proposed method for

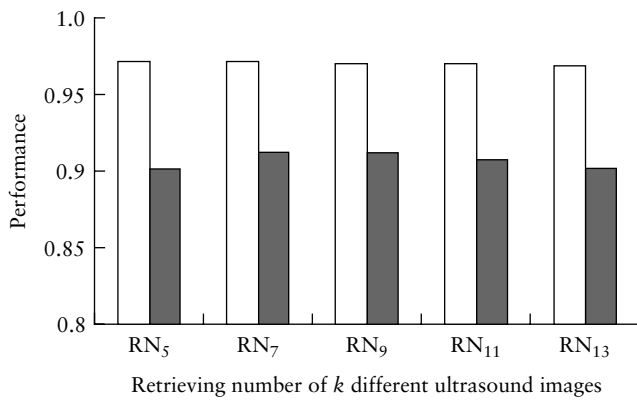


Figure 6 A_Z value (□) and diagnostic accuracy (■) achieved with the proposed computer-aided diagnostic system.

Table 1 The performance of retrieving a number (n) of different ultrasound images (denoted by RN _{n})

Parameter	RN ₅ (%)	RN ₇ (%)	RN ₉ (%)	RN ₁₁ (%)	RN ₁₃ (%)
Accuracy	90.2	91.3	91.2	90.8	90.2
Sensitivity	96.1	96.5	97.0	93.5	96.5
Specificity	86.5	88.1	87.6	89.2	86.2
PPV	81.5	83.5	82.9	84.3	81.3
NPV	97.3	97.6	97.9	95.7	97.6

NPV, negative predictive value; PPV, positive predictive value.

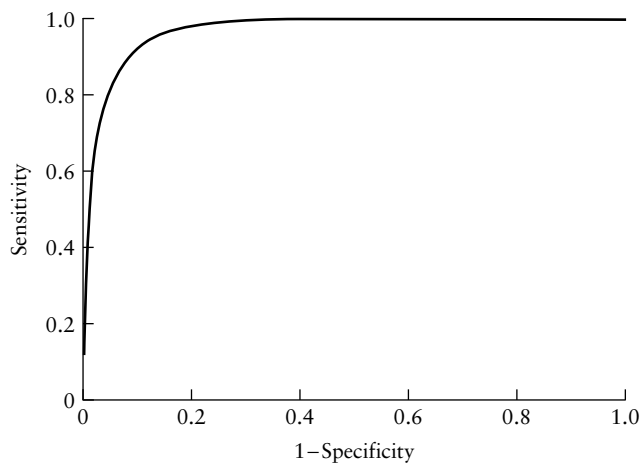


Figure 7 Receiver-operating characteristics (ROC) curve for the retrieval technique employed in classifying malignant and benign tumors (the A_Z value for the ROC curve is 0.970 ± 0.006).

classifying tumors on various ultrasonic systems, the ultrasound image database was divided into four groups based on the ultrasonic system model. The simulation was made as the k -fold cross-validation method. For example, the first group (i.e. all ultrasound images acquired from the SDD 1200 scanner) was excluded and the remaining three groups, images acquired from three other ultrasonic systems, were functioned as the training set. The process was repeated until all four

Table 2 The performance of the retrieval of nine ultrasound images with different threshold values (Th)

Th	True-positives	True-negatives	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 image retrieval technique with $Th = 0.3$ for RN₉

Ultrasound image classification	Benign*	Malignant*
Benign ($DS < Th$)	TN 324	FN 7
Malignant ($DS \geq Th$)	FP 46	TP 223
Total	370	230

*Histological finding. FN, false-negative; FP, false-positive; TN, true-negative; TP, true-positive.

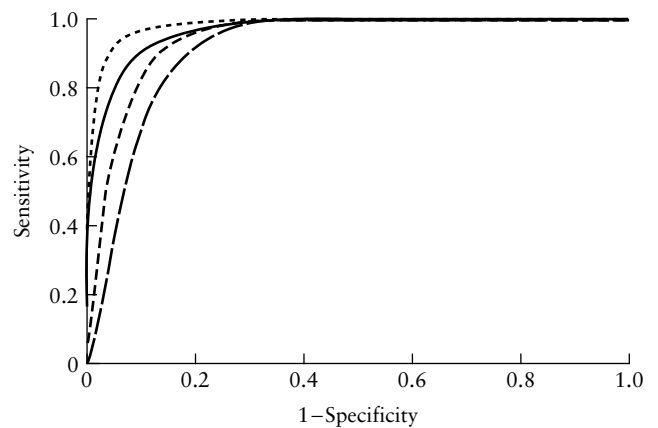


Figure 8 Receiver-operating characteristics curve analysis of all four sets of different ultrasonic systems. - - - - -, A_Z = 0.9434 HDI 3000; - · - · - ·, A_Z = 0.9845 HDI 5000; — — —, A_Z = 0.9129 LOGIQ 700; ———, A_Z = 0.9653 SDD 1200.

groups had functioned in turn as a testing group. The ROC analysis and A_Z values of four sets are presented in Figure 8. As Figure 8 demonstrates, the proposed method an excellent diagnostic performance for all four sets of different ultrasonic systems measured by A_Z value. The average diagnostic time for a breast ultrasound image was less than 5 ms.

DISCUSSION

The significant characteristic of the present study is that the data were obtained from four quite distinct commercial ultrasound systems. Although some machine dependence was still evident, the texture algorithms proved robust enough to permit a clear separation between benign and malignant lesions, independent of the ultrasound scanner recording the data. This is a

reasonably successful result given the nature of the ultrasound scanners employed. The ability of the proposed CAD system to be used with various ultrasound machines is mainly due to the use of preprocessing techniques that homogenize texture features between systems. The PCA technique is employed to obtain a lower dimensional textural vector that reduced the training and diagnosis time dramatically.

The American Cancer Society¹⁸ observed that an accurate and reliable diagnostic procedure is the most significant factor in early diagnosis. Mammography and sonography are frequently employed clinical practices and such modalities can help physicians differentiate benign breast tumors from malignant lesions. Although breast sonography plays a role as an auxiliary to mammography, ultrasound examination is more convenient and safer than mammography for patients undergoing regular physical examination. Controversy exists about the utility of ultrasound images for diagnosing breast cancer, because of the many heterogeneous and overlapping characteristics shared between malignant and benign lesions. Stavros *et al.*¹⁹ indicated that the ultrasound technique is helpful for diagnosing breast cancer more precisely. The authors note that the ultrasound technique needs an accomplished radiologist and extensive real-time evaluation. Physicians use mammography and sonography to diagnose breast cancer via visual experiences²⁰. Physicians with varying experiences might have different interpretations of breast ultrasound images. To avoid needless biopsy and enhance the diagnostic accuracy, a CAD system can provide a second beneficial support reference.

Rapidly developing ultrasound technologies have led to the use of many different ultrasonic systems in medical diagnosis. The main concerns when designing a CAD system for various ultrasonic systems are resolution and contrast. How to transform the information needed for diagnosis between different systems becomes a significant issue. The users care about whether a designed system is suitable for another ultrasonic machine without any amendment or through the adjustment of the parameters using intelligent selection algorithms according to the various ultrasonic machines. Our previous study proposed a novel diagnostic system for various ultrasonic systems in which inter-pixel correlation on the ultrasound images were employed to differentiate benign and malignant tumors. Accessing the information needed between two different systems is achieved through the proposed adjustment technique²¹. However, adjustment schemes for various ultrasonic systems are still necessary. The more ultrasonic systems that exist, the greater the efforts that must be expended to transform the information among them. This transformation is a tedious process, which the proposed new algorithms can improve. The image retrieval technique uses the projected principal vector to query the ultrasound images with similar textures from the database. Consequently, the perplexity training procedure can be avoided. Furthermore, historical cases can be directly added into the reference database without retraining. With the expansion of the database, new cases

can easily be gathered and used as references. Figure 8 reveals that a correct diagnosis may be accomplished by referring to the retrieved images obtained from the various ultrasonic scanners, demonstrating that the high diagnostic accuracy rates were not due to retrieving images from one particular ultrasonic machine but came from randomized retrievals from various ultrasonic systems.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC93-2213-E-029-014.

REFERENCES

1. Chen DR, Chang RF, Huang YL. Breast cancer diagnosis using self-organizing map for sonography. *Ultrasound Med Biol* 2000; **26**: 405–411.
2. Chen DR, Chang RF, Kuo WJ, Chen MC, Huang YL. Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks. *Ultrasound Med Biol* 2002; **28**: 1301–1310.
3. Chen DR, Chang RF, Huang YL. Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. *Radiology* 1999; **213**: 407–412.
4. Chen DR, Chang RF, Huang YL, Chou YH, Tiu CM, Tsai PP. Texture analysis of breast tumors on sonograms. *Semin Ultrasound CT MR* 2000; **21**: 308–316.
5. Garra BS, Krasner BH, Horii SC, Ascher S, Mun SK, Zeman RK. Improving the distinction between benign and malignant breast-lesions – the value of sonographic texture analysis. *Ultrason Imaging* 1993; **15**: 267–285.
6. Haykin S. *Neural Networks: A Comprehensive Foundation* (2nd edn). Prentice Hall: Upper Saddle River, NJ, 1999.
7. Kuo WJ, Chang RF, Lee CC, Moon WK, Chen DR. Retrieval technique for the diagnosis of solid breast tumors on sonogram. *Ultrasound Med Biol* 2002; **28**: 903–909.
8. Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Mach Intell* 1996; **18**: 837–842.
9. Gimelfarb GL, Jain AK. On retrieving textured images from an image database. *Pattern Recognition* 1996; **29**: 1461–1483.
10. Gudivada VN, Jung GS. An architecture for and query processing in distributed content-based image retrieval. *Real-Time Imaging* 1996; **2**: 139–152.
11. Gudivada VN, Raghavan VV. Content-based image retrieval-systems. *Computer* 1995; **28**: 18–22.
12. Maess B, Friederici AD, Damian M, Meyer AS, Levelt WJM. Semantic category interference in overt picture naming: sharpening current density localization by PCA. *J Cogn Neurosci* 2002; **14**: 455–462.
13. Sinha U, Kangarloo H. Principal component analysis for content-based image retrieval. *Radiographics* 2002; **22**: 1271–1289.
14. Costa S, Fiori S. Image compression using principal component neural networks. *Image and Vision Computing* 2001; **19**: 649–668.
15. Gonzalez RC, Woods RE. Image enhancement in the spatial domain. In *Digital Image Processing*. Addison Wesley: Reading, MA, 2002; 75–146.
16. Gonzalez RC, Woods RE. Image compression. In *Digital Image Processing*. Addison Wesley: Reading, MA, 2002; 409–518.
17. Weiss SM, Kapouleas I. An empirical comparison of pattern recognition neural nets and machine learning classification

methods. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*. Morgan Kaufman: San Mateo, CA, 1989; 234–237.

18. *Breast Cancer Facts and Figures 2001–2002*. American Cancer Society: Atlanta, GA, 2003.
19. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid Breast nodules – use of sonography to distinguish benign and malignant lesions. *Radiology* 1995; **196**: 123–134.
20. Skaane P, Engedal K. Analysis of sonographic features in the differentiation of fibroadenoma and invasive ductal carcinoma. *AJR Am J Roentgenol* 1998; **170**: 109–114.
21. Kuo WJ, Chang RF, Moon WK, Lee CC, Chen DR. Computer-aided diagnosis of breast tumors with different US systems. *Acad Radiol* 2002; **9**: 793–799.
22. Jolliffe IT. *Principal Component Analysis*. Springer-Verlag: New York, NY, 1986.

APPENDIX

Auto-covariance coefficients of a tumor region-of-interest (ROI) subimage can be regarded as a feature vector. Assuming that there are N feature vectors in the training set, the average feature vector m from the training set is given by:

$$m = \frac{1}{N} \sum_{i=1}^N \bar{x}_i, \quad (6)$$

where \bar{x}_i is the $\Delta m \times \Delta n$ dimension feature vector corresponding to i th ROI subimage in the training set. An ROI subimage will produce a 49-D textural feature vector (auto-covariance coefficients with Δm and Δn are 7). The value of $\gamma(0, 0)$ is always 1 for a normalized auto-covariance matrix. Excluding the element $\gamma(0, 0)$, other auto-covariance 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 v_n be the eigenvectors of O :

$$O_{N \times N} = \begin{bmatrix} (x_1 - m) \cdot (x_1 - m) & \cdots & (x_1 - m) \cdot (x_N - m) \\ \vdots & \ddots & \vdots \\ (x_N - m) \cdot (x_1 - m) & \cdots & (x_N - m) \cdot (x_N - m) \end{bmatrix}_{N \times N}. \quad (7)$$

These eigenvectors establish linear combinations of the training set to form the basis set of vectors u_i . The best characteristics of the variation in the training vectors can be represented by principal component u_i :

$$u_i = \sum_{k=1}^N v_{ik} (\bar{x}_k - m), \quad (8)$$

for $i = 1, 2, \dots, N$. The basis set vectors associated with the largest eigenvalues capture most of the information about the feature vectors in the training set. The percentage of total variability explained by each principal component can be estimated. Generally, the first p principal components to exceed 90% of the total variance of the original vectors will be used to approximately project the original feature vector x_k into a new p -dimensional feature vector²². The approximation equation is defined as:

$$x_k \approx \sum_p \omega_p \mu_p. \quad (9)$$

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 the coefficients w_q computed.