


私立東海大學
資訊工程與科學研究所

碩士論文
指導教授：黃育仁

乳癌超音波影像電腦輔助診斷之研究

**A Study of Breast Cancer Computer-Aided
Diagnosis on Sonogram**



研究生：劉雅光

中華民國 九十三年 六月

摘要

由於飲食習慣的西化與日趨嚴重的環境污染，乳癌已經成為東方女性死亡的主因。而早期的診斷與治療可有效降低乳癌的致死率，早期的診斷需仰賴精確的檢驗工具，醫學超音波影像在非侵入式的乳癌檢驗中，具有快速與方便的優點，已成為臨床醫師在日常檢查中的利器，但由於超音波影像中常含有大量的雜訊及斑點，缺乏經驗的臨床醫師常有錯誤診斷的情形，因此超音波影像電腦輔助診斷系統(CAD)對於提升診斷的精確度有其必要性。

本論文提出兩個電腦輔助診斷系統，皆利用超音波腫瘤影像中的紋路特徵來分辨腫瘤影像的良惡性，第一個方法是利用 Support vector machine (SVM)來區別腫瘤影像的良惡性；第二個方法則是利用影像搜尋技術來診斷腫瘤。先前的電腦輔助診斷研究常使用類神經網路來判別腫瘤的良惡性，但是類神經網路的訓練程序非常耗時，而且起始參數調整十分複雜，為了改善此一架構的缺點，所以本論文的第一個方法將類神經網路替換成 SVM 來判別腫瘤的良惡性，SVM 最主要的優點為訓練步驟非常快速而且穩定，大大降低了歷史病例訓練與進行診斷所需的時間，並且提高了乳癌診斷的正確性。

然而大部份的電腦輔助診斷系統只針對單一機型的超音波腫瘤影像進行測試，無法直接使用於其他不同機型的超音波儀器，因此本論文的第二個方法利用影像搜尋的技術排除不同機型超音波儀器間的差異，以提高電腦輔助診斷系統的實用性，此系統的主要優點是可以直接將新增病例加入腫瘤影像資料庫中，進而

辨別腫瘤影像的良惡性，不需要進行重新訓練的步驟。本論文中提出兩個電腦輔助診斷系統皆經過實際病例的測試，並且得到令人滿意的效果，希望此研究能造福更多的婦女，達到早期診斷與早期治療的目的。

ABSTRACT

This thesis proposed two computer-aided diagnosis (CAD) systems to identify breast cancer on sonogram. At first the physician located regions-of-interest (ROI) subimage in ultrasound image. The textual features in the ROI subimage were utilized to classify breast tumors. The proposed CAD systems using inter-pixel textual features classified the tumor as benign or malignant. The first CAD system utilized support vector machine (SVM) as classifier in the differential diagnosis of solid breast tumors. The SVM system differentiates solid breast nodules with a relatively high accuracy and helps inexperienced operators avoid misdiagnosis. The main advantage in the proposed SVM system is that the training procedure was very fast and stable. The training and diagnosis procedure of the proposed system is almost 700 times faster than that of multilayer perception neural networks (MLPs). With the growth of the database, new ultrasonic images can be collected and used as reference cases while performing diagnoses. This study reduces the training and diagnosis time dramatically.

Successful applications of the CAD strategy have also been reported for other types of texture analysis. However, most of the strategies performed in a specific US machine and they do not show the result whether the strategy performs well for different US systems. The second CAD system exploited image retrieval techniques

with textural features to classify benign and malignant breast tumors on different US systems. However, the textural features always perform as a high dimensional vector. High dimensional vector is unfavorable to differentiate breast tumors in practice. This thesis employs the principal component analysis (PCA) to project the original textural features into a lower dimensional principal vector that captured most of the textural information. Image retrieval techniques were utilized to differentiate breast tumors based on the similarity of the principal vectors. The query ROI subimages were diagnosed as malignant or benign tumors according to properties of retrieved images from the US image database. In the proposed CAD system, historical cases can be directly added into the database and the retraining procedure is unnecessary. The proposed CAD system differentiates solid breast nodules with a relatively high accuracy in the different ultrasonic systems.

Keyword: computer-aided diagnosis, breast cancer, support vector machine, image retrieval, principal component analysis

ACKNOWLEDGEMENTS

During pursuing studies in the Tunghai University, I wish to express my sincere gratitude to many people.

Firstly, I would like to show my great appreciation to my advisor Dr. Yu-Len Huang, for his painstaking guidance, patient training, and unwearied discussions. Besides, I would also like to thank Dr. Ruey-Feng Chang, for his precious suggestion to this thesis. And I also want to thank Dr. Rung-Ching Chen, who has been giving me some valuable opinions.

Then I wish to thank the colleagues in the Image and Video Processing Laboratory at the Department of Computer Science and Information Engineering Tunghai University, for their helps and friendships during preparing this thesis.

Finally, I am grateful to my family for their support and encouragement that I can concentrate more on my study. I also want to dedicate this thesis to everyone who has assisted me during those times.

TABLE OF CONTENTS

摘要.....	1
ABSTRACT.....	3
ACKNOWLEDGEMENTS.....	5
TABLE OF CONTENTS.....	6
LIST OF FIGURES.....	7
LIST OF TABLES.....	8
CHAPTER 1 INTRODUCTION.....	9
CHAPTER 2 TECHNIQUE REVIEW.....	14
2.1 Support Vector Machine (SVM).....	14
2.2 Principle Component Analysis (PCA).....	16
2.3 Textural Analysis.....	18
CHAPTER 3 SVM FOR BREAST CANCER DIAGNOSIS.....	20
3.1 Breast Cancer Diagnosis.....	20
3.2 Ultrasound image database.....	21
3.3 Experiments and results.....	22
CHAPTER 4 IMAGE RETRIEVAL TECHNIQUE FOR BREAST CANCER DIAGNOSIS.....	29
4.1 Breast Cancer Diagnosis.....	29
4.2 Data Acquisition.....	33
4.3 Image Preprocessing.....	34
4.4 Experiments and results.....	36
CHAPTER5 CONCLUSION AND DISCUSSION.....	43
REFERENCES.....	45

LIST OF FIGURES

Fig. 1. Support vector machine optimal hyperplane.....	16
Fig. 2. (a) A 736×556 full breast sonogram and (b) The ROI rectangle is approximately $2.91\text{cm} \times 2.33\text{ cm}$ in size, captured with a resolution of 169×135 pixels (a $1\text{ cm} \times 1\text{cm}$ rectangle contains 58×58 pixels).....	23
Fig. 3. The diagram of the ROC curve for the SVM in the classification of malignant and benign tumors. The A_z value for the ROC curve is 0.9695 ± 0.0150	25
Fig. 4. The performance of SVM with different parameter γ (gamma) values.....	25
Fig. 5. The bar graph of the first three principal components explain over 95% of the total variability in the standardized ratings.....	30
Fig. 6. Flow chart of the proposed CAD system.....	32
Fig. 7. ROI subimages of breast lesions acquired by (a) SDD 1200 scanner, (b) HDI 3000 scanner, (c) HDI 5000 scanner and (d) LOGIQ 700 scanner	36
Fig. 8. (a) An original ROI subimage, (b) the preprocessed image, (c) histogram of original ROI subimage and (d) histogram of the preprocessed ROI	37
Fig. 9. The bar graph of the A_z value and diagnostic accuracy achieved with the proposed CAD system (error bars indicate 1 SD)	38
Fig. 10. 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 0.970 ± 0.006).....	39
Fig. 11. The bar graph of the percentage of images of specific ultrasonic system in the entire database (denotes PS) and the Percentage of retrieved and query images from the identical ultrasonic system (denotes PR)	41
Fig. 12. The ROC analysis of all four sets of different ultrasonic systems	42

LIST OF TABLES

Table 1. The number of misdiagnosed cases of the MLPCAD and the proposed SVMCAD for each test set.	27
Table 2. Classification of breast nodules by proposed SVM system with $\gamma = 0.01$	27
Table 3. The performance of the proposed SVM system with $\gamma = 0.01$	28
Table 4. The computation time by using proposed SVMCAD and the MLPCAD.....	28
Table 5. The performance of retrieving number of k different US images (denote RN_k)	39
Table 6. The performance of the retrieval nine US images are with different threshold values for RN_9	40
Table 7. Classification of Breast nodules by proposed image retrieval technique with $Th = 0.3$ for RN_9	41

CHAPTER 1

INTRODUCTION

Breast cancer is one of the leading causes of deaths for female population in both the east and the west. Although westerner is easier occurred breast cancer in the early period, easterner increases a rapid of cancer cases recently because of the changes of lifestyle and environment. The disease has come into public notice. American Cancer Society (ACS) [1] indicated that accurate and reliable diagnostic procedure are the most influential in the early diagnosis.

Mammography and ultrasonography are the most frequently adopted medical imaging in clinical practices for patients. Those modalities can help physicians to differentiate benign breast tumors from malignant lesions. Ultrasound (US) is usually an auxiliary for mammography to diagnose tumor, but the US inspection is more convenient and safer than mammography in the routine physical examination. The controversy on the utility of sonogram for differentiating breast cancer is there are heterogeneous and much overlap characteristics between malignant and benign lesions. Stavros et al. [2] pointed out the sonogram technique is a useful tool to help physicians diagnosing breast cancer more accurate. The authors indicated the US technique required a practiced radiologist with extensive real-time evaluation, because physicians with different experiences via visual experiences [3] might give different

interpretations of breast sonograms. Biopsy is expensive and baneful for patients because of a great quantity of indeterminate lesions need to be differentiated. In order to avoid unnecessary biopsy and improve the accurate of diagnosis, a reliable computer-aided diagnosis (CAD) system is demanded for a physician to support as a second beneficial reference. This study proposed two effective CAD systems to diagnose breast cancer.

Chen et al. [4-6] utilized the textural features in breast sonogram and neural network (NN) classifiers to differentiate between benign and malignant tumors. The textural variation in the US image has been found as a beneficial feature to identify benign and malignant tumours [7,8]. Chen's CAD systems utilized the multilayer perception neural network (MLP) to classify breast tumors which performed a good diagnostic result. However, the MLP learning procedure is very time-consuming and the diagnostic performance usually depend on initial parameter setting [9], i.e. number of neurons, learning rate, and moment value are hard to decide. The selections of initial parameters unfortunately will affect the results. Contrary to the defects of MLP, the support vector machine (SVM) is feasibility and superiority to extract higher-order statistics to differentiate breast tumors. Thus this study employs the SVM model as a classifier instead of MLP model to differentiate the malignant from benign breast tumor. The diagnostic result of using SVM is fast and steady in breast

tumors classification, thus it is a reliable choice for the proposed CAD system.

The amount of digital images recently has tremendously increased for image database applications, for example, digital libraries, picture archiving and communications systems (PACS), geographic information systems (GIS), and etc. Although most applications of the computer-aided diagnosis (CAD) strategies perform well in a specific ultrasonic machine successful, they do not show if the result of the strategies performs well for different ultrasonic systems. The second CAD system of this study used image retrieval technique to distinguish malignant from benign masses of the breast sonogram from different sonogram systems. Characteristics of different types of images is the goal that effective content-based image retrieval technique has to aim at [10,11]. The development of content-based image queries and retrievals are essential to find the desired images automatically from the image database [12,13] while image databases contain a large number of images. Several content-based image retrieval systems, like IBM's QBIC project, provide image retrieval capability to automatic index and query image.

Kuo et al. [14] developed an image-retrieval CAD system to diagnose breast cancer on sonogram. Their system performed a satisfied result for diagnosing breast cancer, but it required extensive evaluation for the weighting coefficients of the feature parameters. Therefore this thesis proposed the second CAD system using

image retrieval to differentiate varieties of breast tumors. Besides, the proposed CAD system utilized auto-covariance matrix as textural feature to identify breast tumour. However, the textual feature vector is always in a high dimensional space. The system is incapable of utilizing the feature vector to identify breast tumors by image retrieval. Thus this study perform the principal component analysis (PCA) [15,16] to diminish the dimension of the feature vector. The original textural feature vector will be mapping into principal vector with a lower dimension [17]. The projected vector called principal vector is used as new textural features 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 ultrasonic image. This study utilized image retrieval techniques to differentiate breast tumors with a relatively high accuracy on differential sonogram systems. With the growth of the database, more and more confirmable breast US images may be collected to be referable cases while diagnosing. Retraining procedure is unneeded in the proposed system and historical cases can be directly added into the database.

Chapter 2 introduces the background of SVM, PCA, and textural features that used in this thesis, Chapter 3 describes the step of using SVM to diagnose breast cancer is explained and experimental results. The step of using image retrieval technique to diagnose breast cancer and experimental results are presented in Chapter

4. Finally, the conclusion and discussion were drawn in Chapter 5.

CHAPTER 2

TECHNIQUE REVIEW

The first CAD system in this thesis used the SVM to differentiate breast tumors. The second CAD system utilized image retrieval technique to discriminate breast tumors. In this chapter, we review techniques appeared in this thesis, including SVM model, PCA method and the modified auto-covariance coefficients.

2.1 Support Vector Machine (SVM)

SVM is aimed at inventing a computationally efficient way of learning separating hyperplanes in a high dimensional feature space [18,19]. The SVM models have been applied for many real-world problems to be an efficient method because of its high generalization performance without necessarily adding a priori knowledge. Thus SVMs recently have been utilized as a useful tool for image recognition, such as hand-written digit recognition, and bioinformatics [20-25].

The SVM can map the input vectors into a high dimensional feature space through some nonlinear mapping which was previously chosen. An optimal separating hyperplane is constructed in this space. The SVM is generally an implementation of the structural risk minimization principle. Object of the principle is to minimize the upper bound on the generalization error. Given a set of training vectors (l in total)

belonging to separate classes, $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$, where $x_i \in R^n$ denotes the i^{th} input vector and $y_i \in \{+1, -1\}$ is the corresponding desired output.

The maximal margin classifier aims to find a hyperplane $w: wx + b = 0$ to separate the training data. Only one maximizes the margin (distance between the hyperplane) and the nearest data point of each class in the possible hyperplanes. Figure 1 shows the optimal separating hyperplane with the largest margin. The support vectors denote the points lying on the margin border. The solution to the classification is given by the decision function

$$f(x) = \text{sign} \left(\sum_{i=1}^{N_{SV}} \alpha_i y_i k(s_i, x) + b \right), \quad (1)$$

where α_i is the positive Lagrange multiplier, s_i are the support vectors (N_{SV} in total), and $k(s_i, x)$ is the function for convolution of the kernel of the decision function.

Three typical kernel functions can be applied is described as below.

Polynomial kernels:

$$k(x, y) = (x \cdot y + 1)^d, \quad (2)$$

where $d \in N$ denotes the degree of the polynomial decision surface.

Radial kernels:

$$k(x, y) = \exp \left(-\gamma (x - y)^2 \right), \quad (3)$$

where $\gamma \in R$ is a non-zero parameter.

Multilayer perception kernels:

$$k(x, y) = \tanh(ax \cdot y + b), \quad (4)$$

where a and b are the parameter scale and offset for the Multilayer perception.

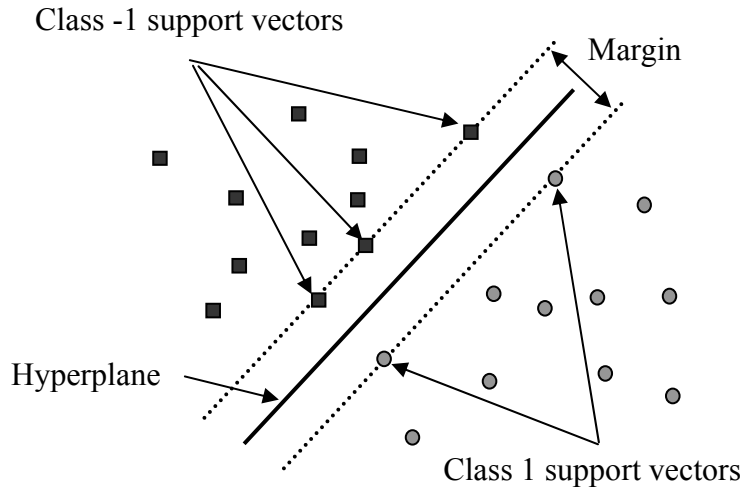


Fig. 1. Support vector machine optimal hyperplane

2.2 Principle Component Analysis (PCA)

The idea behind the PCA is to find another more applicable representation for diminishing the dimension of the original vectors. PCA is a well-known statistical processing technique that provides for reducing redundancy by projecting the original data over a proper basis. Assume that there are N input vectors in the training set. The average vector m from the training set is given by

$$m = \frac{1}{N} \sum_{i=1}^N \bar{x}_i, \quad (5)$$

where \bar{x}_i is the input vector representing the i th image in the training set. 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 is the eigenvectors of O .

$$O_{N \times N} = \begin{bmatrix} (x_1 - m) \cdot (x_1 - m) & \cdots & (x_1 - m) \cdot (x_N - m) \\ \vdots & \ddots & \vdots \\ (x_N - m) \cdot (x_1 - m) & \cdots & (x_N - m) \cdot (x_N - m) \end{bmatrix}_{N \times N}. \quad (6)$$

These eigenvectors determine 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 :

$$u_i = \sum_{k=1}^N v_{ik} (\bar{x}_k - m), \quad (7)$$

for $i = 1, 2, \dots, N$. The basis set vectors associated with the largest eigen-values capture most of the information of the input vectors in the training set. The percent of the total variability explained by each principal component can be calculated. Usually, the first p principal components exceed to 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 vector [26]. The approximation equation is defined as

$$x_k \approx \sum_p \omega_p \mu_p. \quad (8)$$

The coefficients w_p are the new feature vector representing the x_k . The textural

feature vector from a query image, 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 image database. The match image is selected from training set depend on the minimum Euclidean distance comparing with the coefficients of the query image.

2.3 Textural Analysis

Different tissues in a US image always have significantly different textures. The textural information extracted from US image has been found to be an efficient feature to classify breast tumors [7-8, 27]. This thesis used the correlation between neighboring pixels within the images as features. The modified auto-covariance coefficients between pixel (i, j) and pixel $(i+\Delta m, j+\Delta n)$ in an image with size $M \times N$ is defined as

$$\gamma (\Delta m , \Delta n) = \frac{A (\Delta m , \Delta n)}{A (0 , 0)}, \quad (9)$$

where

$$A(\Delta m, \Delta n) = \frac{1}{(M-\Delta m)(N-\Delta n)} \sum_{x=0}^{M-1-\Delta m} \sum_{y=0}^{N-1-\Delta n} |(f(x, y) - \bar{f})(f(x+\Delta m, y+\Delta n) - \bar{f})|, \quad (10)$$

where \bar{f} is the mean value of $f(x, y)$. There will form an auto-covariance matrix which size is $\Delta m \times \Delta n$. This thesis utilized the auto-covariance coefficients as the

inter-pixel features to distinguish the differences between benign and malignant tumors.

CHAPTER 3

SVM FOR BREAST CANCER DIAGNOSIS

In this chapter, the step of using SVM to diagnose breast cancer is described and experimental results will be showed. The area A_z under the ROC curve is an index of quantitative measure of the overall performance of a diagnostic system. The A_z value for the ROC curve of the SVMCAD system can achieve 0.9695 ± 0.0150 .

3.1 Breast Cancer Diagnosis

The feature vector for the input of the SVM classifier is the modified normalized auto-correlation matrix. Both Δm and Δn are chosen 5 in this work, so processing a region-of-interest (ROI) sonogram produces a 5×5 auto-correlation matrix (25 auto-correlation coefficients). Because the value of $(0, 0)$ is always 1, thus except for the element $(0, 0)$, other coefficients are formed as a 24-D textural feature vector. Take note of that the output value of the SVM is either -1 or 1. When the output value of a US breast image is larger than 0, the system will classify the tumor in the image as malignancy. Otherwise, when the output value is smaller than 0, the tumor will be diagnosed as benignancy.

3.2 Ultrasound image database

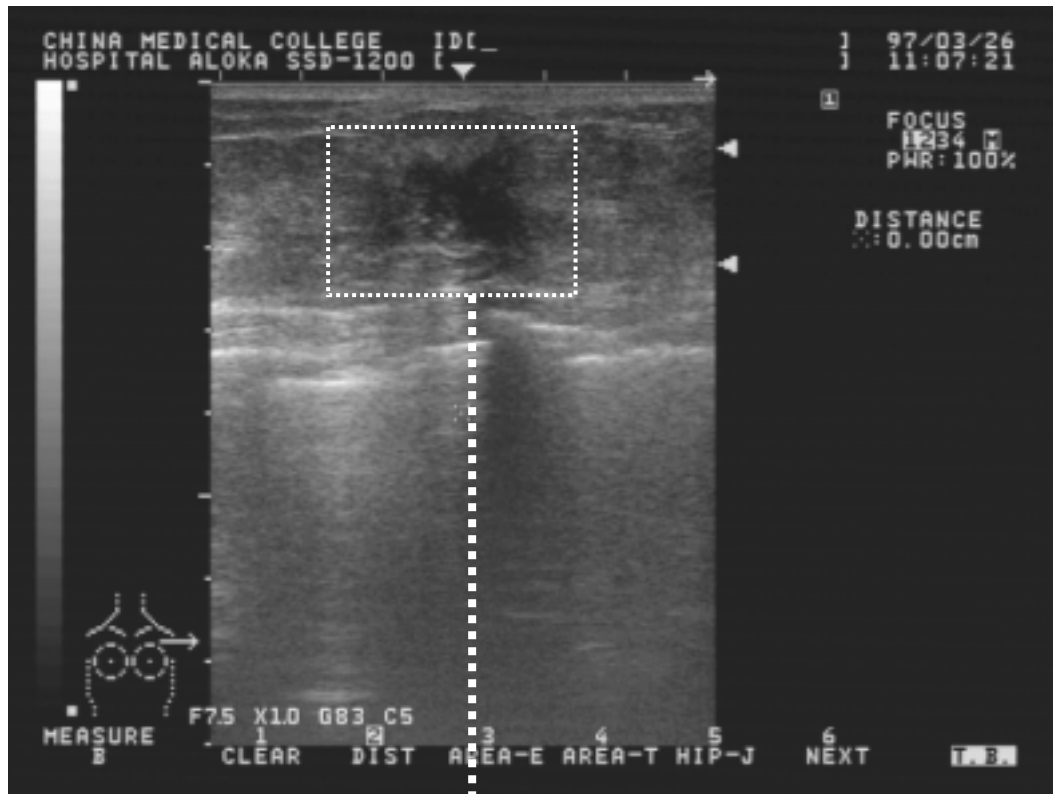
This study supposed that the tumor has already identified by physician. Intensity variation and inter-pixel texture information from the US image were be used to diagnose the tumor. Firstly, physician extracted the rectangular subimage of ROI, and then the computer analyzed textural of the ROI to make a differential diagnosis.

In order to compare the performance with the CAD proposed by Chen et al. [9], the identical US image database is used in this study. The sonogram database consists of 140 images of pathologically proven benign breast tumors from 88 patients, and carcinomas from 52 patients (tumor size > 1 cm in all cases). The database included only one image from each patient. The US images were captured at the largest diameter of each tumor. The images were collected from January 1, 1997 to May 31, 1998; the patients' ages ranged from 18 to 62 years. US image was performed using an ALOKA SSD 1200 (Tokyo, Japan) scanner and with freeze-frame capability and 7.5 MHz linear transducer. No acoustic standoff pad was used in any of the cases. The whole database was supplied by Dr. Chen from Department of General Surgery, China Medical College & Hospital, Taichung, Taiwan. When a sonogram was performed, an analog video signal was transmitted from the VCR output of the scanner to the image-acquiring computer; the data were then digitized using a frame-grabber, Video CATcher (from Top Solution Technology Co.). The capturing

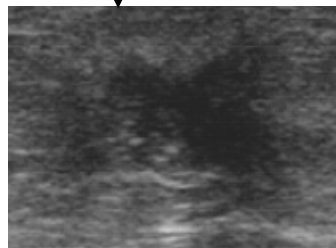
resolution of the external frame grabber was 736×566 pixels for an NTSC video screen picture. Prolab's ProImage software package was used to perform digital sonography in real time. Each monochrome ultrasonic image was quantized into eight bits with 256 gray levels. Dr. Chen manually determined the ROI of the tumor and then saved them in files. Figure 2(a) illustrates a real-time digitized monochrome US image. Figure 2(b) presents an ROI for the tumor.

3.3 Experiments and results

The k-fold cross-validation method [27] and the receiver operating characteristic (ROC) curves (software package LABROC1 by Professor C. E. Metz, University of Chicago) are used to estimate the performance of the proposed SVM system and the MLP system proposed by Chen et al. The k-fold cross-validation method is described as following. The first group is set aside and the remaining $(k - 1)$ groups are used as the training set. Then we set the second group as a testing group and the remaining groups are trained.



(a)



ROI

(b)

Fig. 2. (a) A 736×556 full breast sonogram and (b) The ROI rectangle is approximately $2.91\text{cm} \times 2.33\text{cm}$ in size, captured with a resolution of 169×135 pixels (a $1\text{cm} \times 1\text{cm}$ rectangle contains 58×58 pixels).

This process is repeated until all k groups have been set in turn as a testing group.

We choose k as 10 in the simulations and each group has 14 US images. Two performance measures were used to estimate the performance of the diagnostic system. One measure was the 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 characteristic (ROC) curves. The area A_Z under the ROC curve is an index of quantitative measure of the overall performance of a diagnostic system. To compare performance with different methods could be in accordance with A_Z value.

In this study, the training data and the test data are equivalent to that of the study by Chen et al. [9]. We denoted the proposed SVM system by SVMCAD and the MLP system proposed by Chen et al. by MLPCAD in this thesis. Figure 3 illustrates the diagram of the ROC curve for the SVMCAD in the classification of malignant and benign tumors. If the A_Z value nears to 1 which one could clearly distinguish positive and negative of breast tumors.

Because the performance of using the radial kernels is best in our experimental comparison, hence diagnosis system chooses the radial kernels as kernel function. The overall performance of a diagnostic system usually can be evaluated by examining the ROC area index, A_Z , over the testing output values. The SVMCAD achieve $A_Z = 0.9695 \pm 0.0150$ and MLPCAD achieve $A_Z = 0.9560 \pm 0.0183$, respectively. Moreover, because the radial kernels perform best in the experimental results, the kernels are chosen in the proposed SVM diagnosis system. Figure 4 shows the diagnosis performance for the SVM system with different γ values. With the γ ranged from 0.01

to 0.02, the SVMCAD obtained a stable and the highest accuracy. The system correctly identifies all of the malignant tumors and 81 of 88 benign tumors.

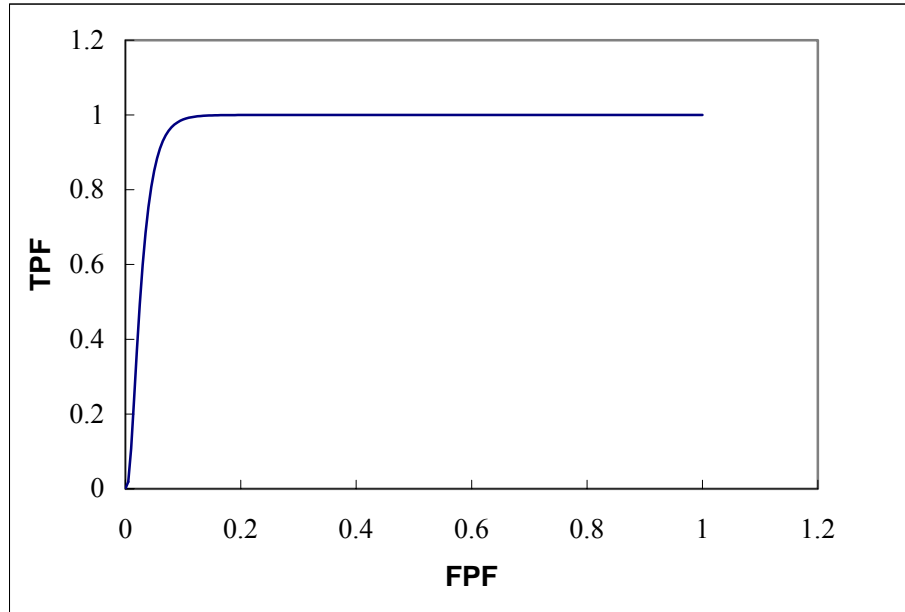


Fig. 3. The diagram of the ROC curve for the SVM in the classification of malignant and benign tumors. The A_z value for the ROC curve is 0.9695 ± 0.0150 .

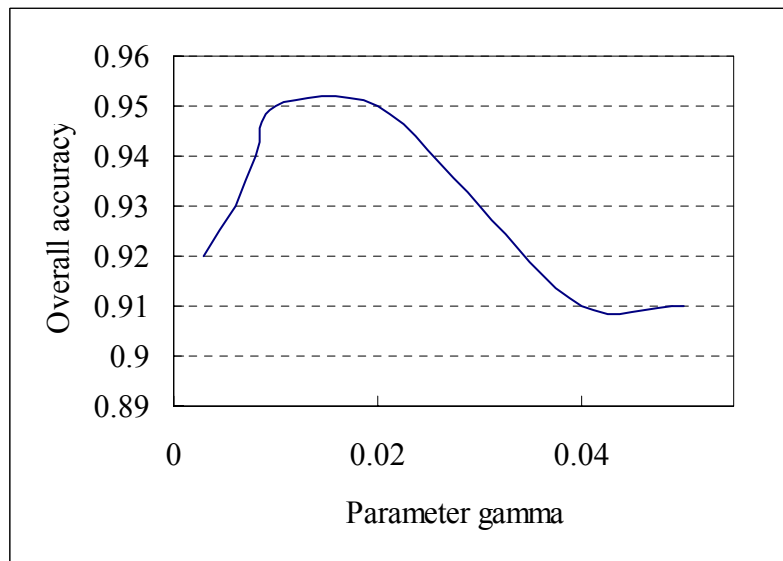


Fig. 4. The performance of SVM with different parameter γ (gamma) values.

Table 1 lists the number of misdiagnosed cases of the MLPCAD (threshold = 0.2) and the SVMCAD for each test set. The accuracy of the proposed SVMCAD for malignancy is 95.0% (133/140), the sensitivity is 98.1% (51/52), the specificity is 93.2% (82/88), the positive predictive value is 89.5% (51/57), and the negative predictive value is 98.8% (82/83), as illustrated in Table 2 and Table 3. On the other hand, the MLPCAD performs that the overall accuracy is 95.0%, the sensitivity of is 98.1%, the specificity is 93.2%, the positive predictive value is 89.5%, and the negative predictive value is 98.8. The computation time for the SVMCAD and the MLPCAD was compared. All simulations are made on a single-CPU Intel PentiumIII-1GHz personal computer with the Windows XP operating system. Table 4 shows the training time for the US image database and the average diagnosis time for each tumour US images. The training and average diagnosis time of the MLPCAD is 700 and 2380 times longer and than that of the proposed SVMCAD, respectively. In all simulation results, the proposed SVMCAD obtains the better classification performance and the speedy computation than those obtained with the MLPCAD.

Table 1. The number of misdiagnosed cases of the MLPCAD and the proposed SVMCAD for each test set

Test Set	MLPCAD (threshold = 0.2)		SVMCAD	
	Malignant Cases	Benign Cases	Malignant Cases	Benign Cases
1	0/5	0/9	0/5	0/9
2	1/5	0/9	0/5	0/9
3	0/5	1/9	0/5	1/9
4	0/5	0/9	0/5	1/9
5	0/5	1/9	0/5	1/9
6	0/5	3/9	0/5	3/9
7	0/5	0/9	0/5	0/9
8	0/5	0/9	0/5	0/9
9	0/6	0/8	0/6	0/8
10	0/6	1/8	0/6	1/8

Table 2. Classification of breast nodules by proposed SVM system with $\gamma = 0.01$

	Benign		Malignant	
SVM Output < 0	TN	81	FN	0
SVM Output ≥ 0	FP	7	TP	52
Total		88		52

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

Table 3. The performance of the proposed SVM system with $\gamma = 0.01$

Item	Proposed SVM system
Accuracy (%)	95.00
Sensitivity (%)	100.0
Specificity (%)	92.05
Positive predictive value (%)	88.14
Negative predictive value (%)	100.0

Accuracy = $(TP+TN)/(TP+TN+FP+FN)$; Sensitivity = $TP/(TP+FN)$; Specificity = $TN/(TN+FP)$; Positive predictive value = $TP/(TP+FP)$; Negative predictive value = $TN/(TN+FN)$.

Table 4. The computation time by using proposed SVMCAD and the MLPCAD.

	Training time (in seconds)	Diagnose time (in millisecond)
MLPCAD	140.71	357.12
SVMCAD	0.20	0.15

The training time is evaluated by using the US image database that contains 140 images. The diagnosis time is average computation time for a US image.

CHAPTER 4

IMAGE RETRIEVAL TECHNIQUE FOR BREAST CANCER DIAGNOSIS

The step of using image retrieval technique to diagnose breast cancer and experimental results are presented in this chapter.

4.1 Breast Cancer Diagnosis

The second CAD system also used the correlation between neighboring pixels within the images as features. However, the textual feature vector catching the various textures in images is always in a high dimensional space. It is detrimental directly performing the high dimensional vector to identify breast tumors. Thus this study used PCA method to reduce the dimension of the feature vector. The method projects original feature vector into a lower dimensional principal vector. The principal vector was performed as the new textural features with a lower dimension to distinguish the differences between benign and malignant tumors. Figure 5 shows the first three principal components ($p = 3$) explain over 95% of the total variability in the standardized ratings. In this study, we found experimentally that $p = 3$ is a good general-purpose choice, so each original 49-D textural feature vector was reduced by PCA method into a new 3-D feature vector.

The common way to find the most similar images from the image database to the new query image is defined in terms of the Euclidean distance of the coefficients w_q and w_p . The retrieved images will be selected from the image database depending on criterion of the Euclidean distance. The proposed CAD system retrieves the first L tumor US images with the smallest Euclidean distances from the US image database.

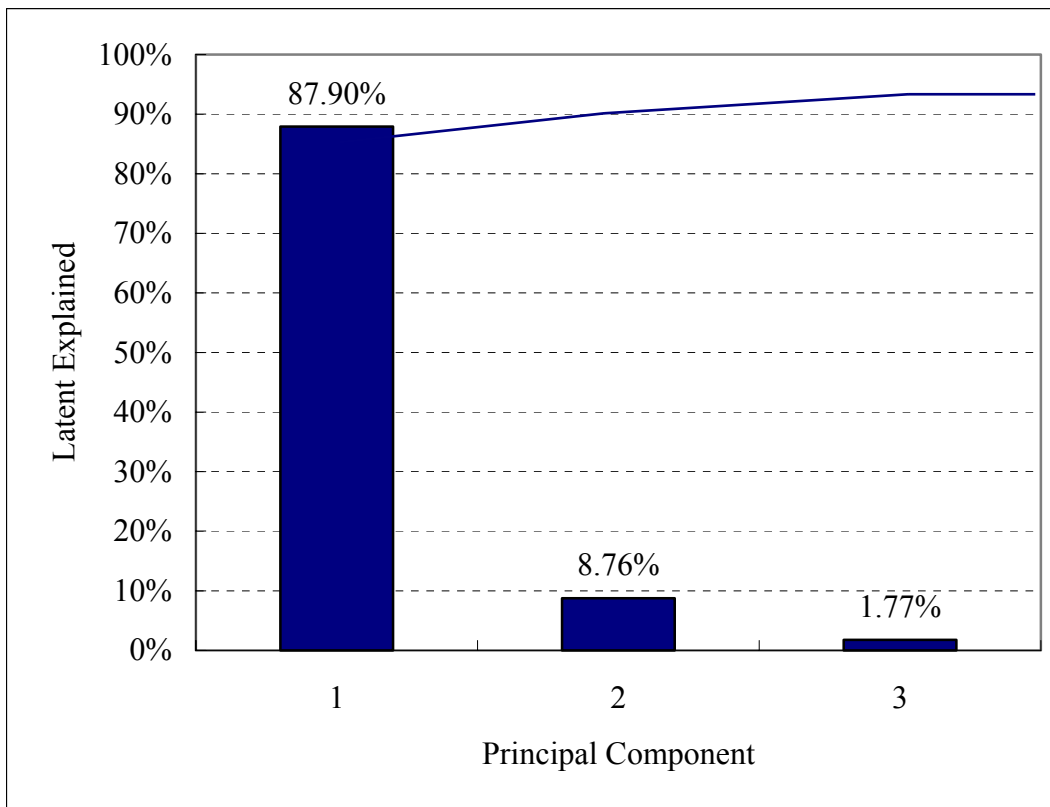


Fig. 5. The bar graph of the first three principal components explain over 95% of the total variability in the standardized ratings

DS value is depending on those retrieved images, the new query image would be diagnosed as benign or malignant lesion. The DS value is defined as

$$DS = \sum_{i=1}^L Weight_i \times Tumor_class_i, \quad (8)$$

$$Weight_i = \frac{L-i+1}{\sum_{j=1}^L j}, \quad (9)$$

$$Tumor_class_i = \begin{cases} 1, & \text{if the retrieved image } i \text{ is malignant case} \\ 0, & \text{if the retrieved image } i \text{ is benign case} \end{cases} . \quad (10)$$

Each retrieved image will be assigned a weight value have to notice. The weight value is determined by the corresponding retrieved order. A cut-off threshold Th is predefined to be a demarcation line to separate breast tumors. If the evaluated DS value is larger than Th , the tumor is classified as malignant tumor. Conversely, if the evaluated DS value is less than Th , the tumor is classified as benign tumor. The flow chart of the proposed diagnostic method is shown in Fig. 6.

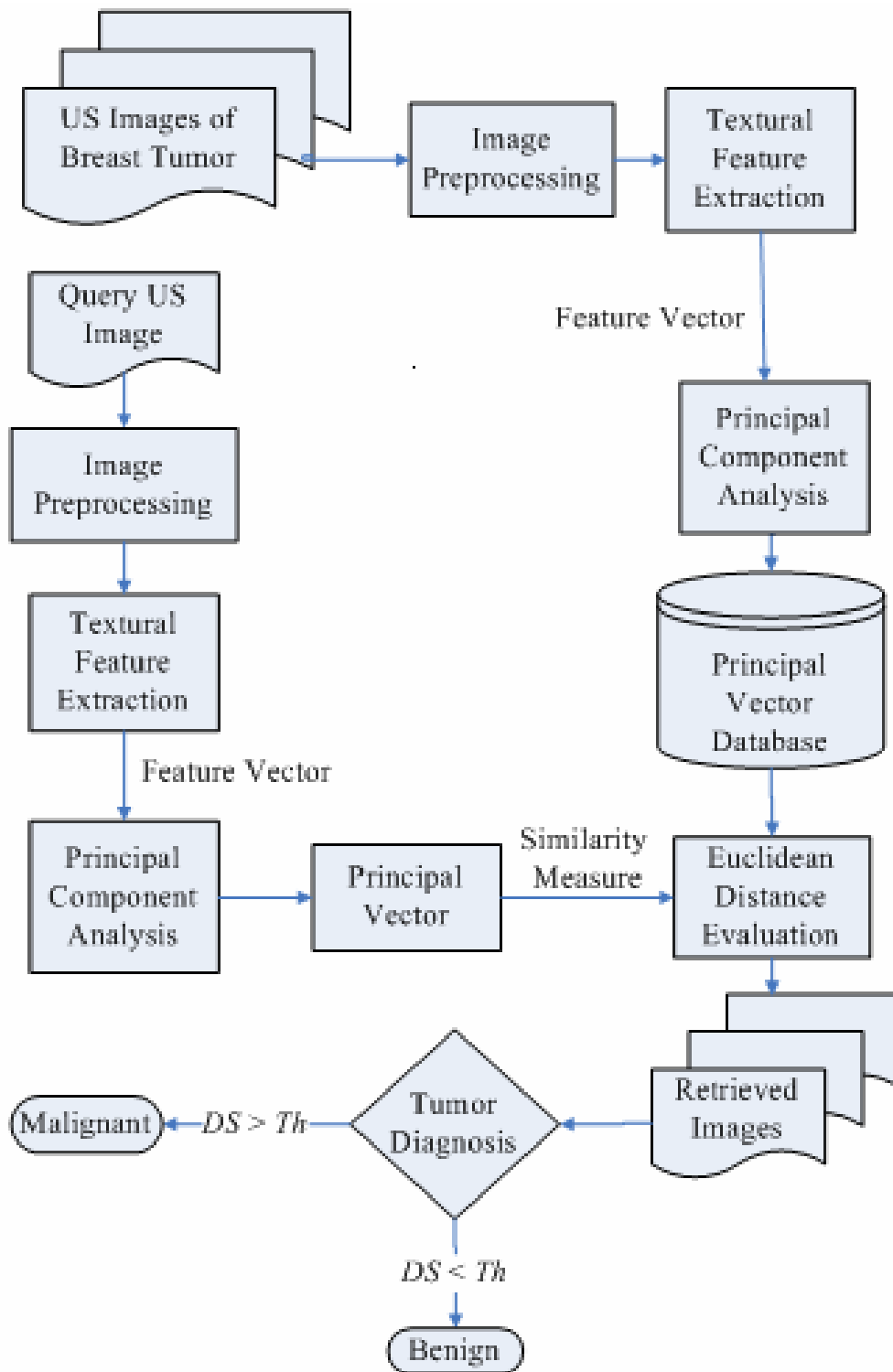


Fig. 6. Flow chart of the proposed CAD system

4.2 Data Acquisition

There are total 600 images of pathologically proven benign breast tumors from 370 patients, and carcinomas from 230 patients (tumor size > 0.8 cm in all cases) in the sonogram database. The sonograms were captured at the largest diameter of the tumor. The images were collected from August 1, 1999 to May 31, 2001; the patients' ages ranged from 17 to 64 years. The breast US images were acquired from the four different US systems:

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

The analog signals obtained with SDD 1200 system from the VCR output of the scanner were transmitted to a frame grabber, Video CATcher (Top Solution Technology, Taipei, Taiwan), and then each monochrome US image was quantized into eight bits with 256 gray levels. Besides, the digital images were obtained prior to

biopsy using by the HDI 3000 system with a 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 US scanning, no acoustic standoff pad was used. Most of the cases are tissue proved and some were followed up at least 2 years. Each image in the US database only extracted from one patient. No acoustic stand-off pad was used. All the images were supplied by Dr. Moon and Dr. Chen. Dr. Moon is from Department of Diagnostic Radiology College of Medicine, Seoul National University Hospital, Seoul, South Korea. The ROI which contains the tumor was selected by Dr. Chen. Through out this study, only the ROI sub-images are used to investigate the texture characteristics of benign and malignant lesions. The physician used the software package, ProImage (Prolab, Taipei, Taiwan), to select the rectangular ROI subimage manually and then saved them in files for textural analysis in real time. The breast US image database utilized the ROI subimages to investigate differential textural characteristic between benign and malignant breast tumor.

4.3 Image Preprocessing

Images acquired from different US systems may lead to influence of distinct gray level contrast. As an example, Fig. 7 shows the ROI subimages of breast tumor acquired from different US systems. The images from the different sonogram systems

are existing variations in contrast clearly. In order to obtain similar contrast for the images in the US image database, a preprocessing adjustment was performed to US images before analyzing the textural feature of the ROI subimages. Histogram manipulation is an effectively way in image processing for image enhancement. Histogram equalization [28] is a mathematical process that could enhance the contrast in the image and reduce variation between images from different ultrasonic systems. Thus all images in the US image database were adopted histogram equalization as preprocessing procedure. The histogram equalization is a good approach because this technique automatically enhances digital image and the results from this technique are predictable. Figure 8 shows the ROI subimages of the original US image, the equalized image and their corresponding histogram distributed figures.

The proposed CAD systems aim at exploiting the correlation between neighboring pixels within the images as features to classify breast tumor. The 2-D normalized auto-correlation coefficients [29] was used to reflect the inter-pixel correlation within an image. The auto-covariance coefficients were used as the feature vector for representing each tumor ROI subimage. In this study, auto-covariance coefficients with Δm and Δn are chosen as 7. Thus every ROI subimage will produce a 49-D textural feature vector. Excluding the element $\gamma(0, 0)$, other auto-covariance coefficients are formed as a 48-D textural feature vector.

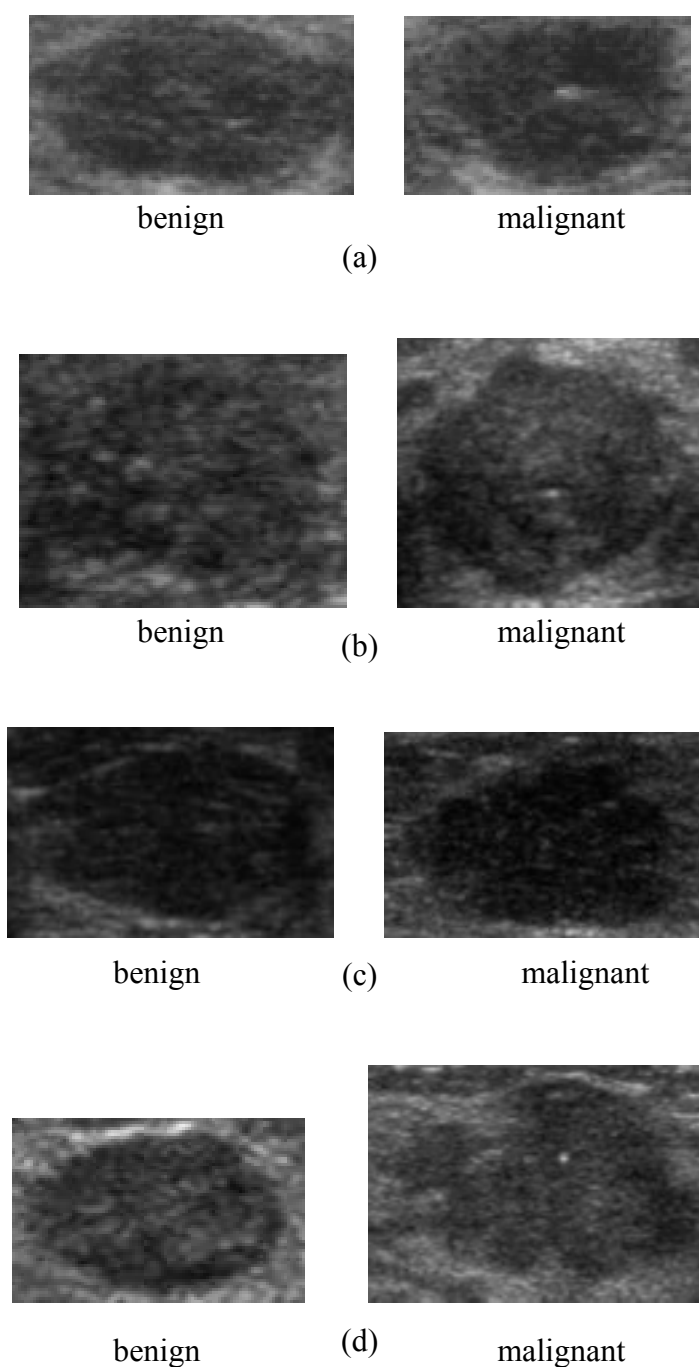


Fig. 7. ROI subimages of breast lesions acquired by (a) SDD 1200 scanner, (b) HDI 3000 scanner, (c) HDI 5000 scanner and (d) LOGIQ 700 scanner

4.4 Experiments and results

The proposed CAD system using retrieval technique also used the k -fold cross-validation method [27] to estimate the performance. The database contains total

600 US images which is divided into k groups in randomly. The k is selected as 10 in the simulations and each group has 60 US images.

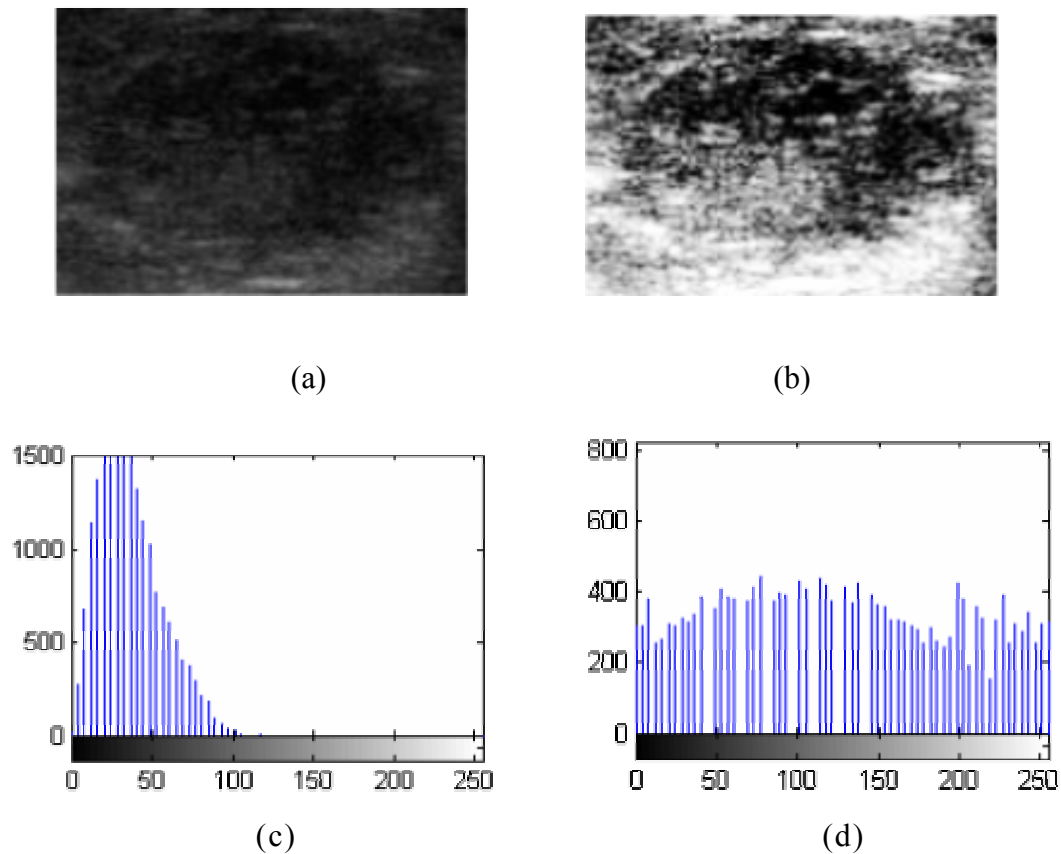


Fig. 8. (a) An original ROI subimage, (b) the preprocessed image, (c) histogram of original ROI subimage and (d) histogram of the preprocessed ROI

In this study, RN_k denotes that the number of retrieving k different US images from the US image database for diagnosing breast tumors. Figure 9 shows that the bar graph of the A_z value and diagnostic accuracy achieved with the proposed CAD system ($k = 5, 7, 9, 11$ and 13). We can find that the performance of retrieving

different numbers of tumor US images were similar. Table 5 shows the comparison of retrieving different numbers of tumor US images to differentiate benign and malignant ones. The RN₉ set achieves good performance in average. Figure 10 illustrates the diagram of the ROC curve for the proposed CAD system. The proposed system with RN₉ achieves $A_z = 0.970 \pm 0.006$.

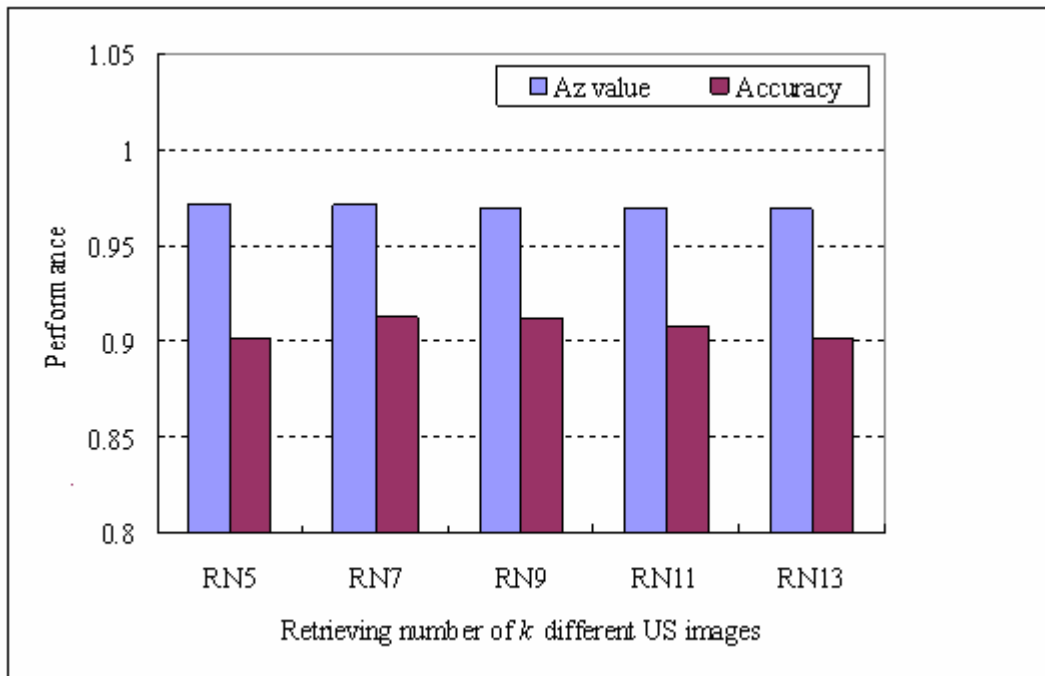


Fig. 9. The bar graph of the A_z value and diagnostic accuracy achieved with the proposed CAD system (error bars indicate 1 SD)

Table 5. The performance of retrieving number of k different US images (denote RN_k)

	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%

Accuracy = $(TP+TN)/(TP+TN+FP+FN)$; Sensitivity = $TP/(TP+ FN)$; Specificity = $TN/(TN+FP)$; PPV = $TP/(TP+FP)$; NPV= $TN/(TN+FN)$

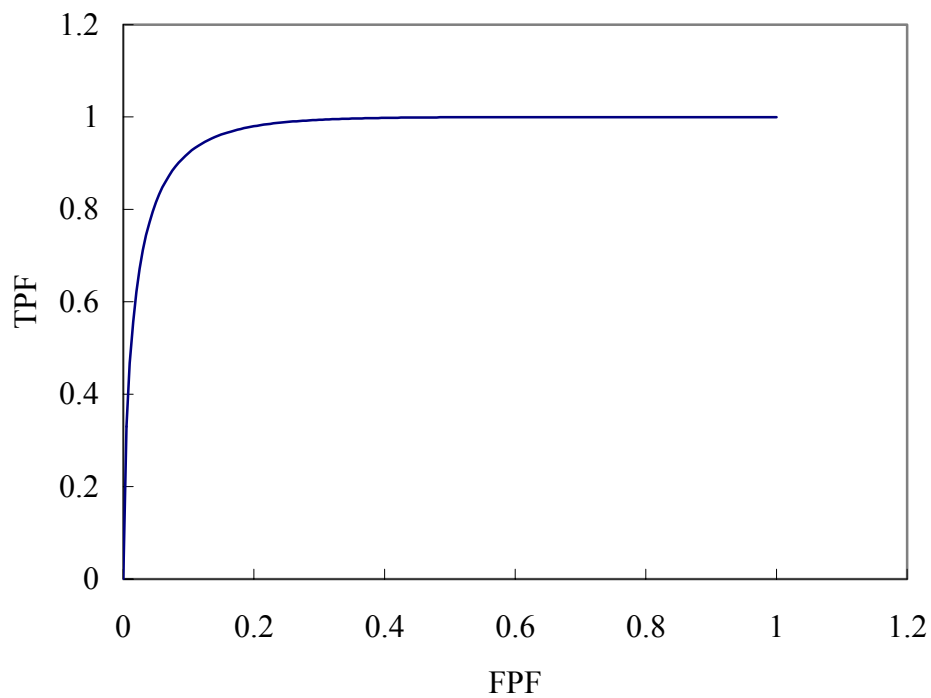


Fig. 10. The diagram of the ROC curve for the retrieval technique is employed in classifying of malignant and benign tumors (the A_z value for the ROC curve is 0.970 ± 0.006)

Table 6. The performance of the retrieval nine US images are with different threshold values for RN_9

Th	True Positive	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 7. Classification of Breast nodules by proposed image retrieval technique with $Th = 0.3$ for RN_9

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

*Histological finding: TP = true-positive; TN = true-negative; FP = false-positive; FN = false-negative

Table 6 makes a comparison with different threshold cut-off values for RN_9 . We found that the ideal threshold cut-off value Th of 0.3 was the better choice. Table 7 summaries the diagnostic performance for RN_9 . The bar graph in Fig. 11 illustrates that the percentage of images of specific ultrasonic system in the entire database, denotes PS, and the percentage of retrieved and query images from the identical ultrasonic system, denotes PR. Clearly, the PR is getting on for PS with the different

ultrasonic systems.

Besides, to prove that the proposed method is practical in classifying tumors on different ultrasonic systems, we also divided the US image database into 4 groups based on the model of ultrasonic system. The simulation is made as the k -fold cross-validation method. For example, the first group, i.e. all US images acquired from SDD 1200 scanner, is set aside and the remaining 3 groups, images acquired from another three ultrasonic systems, are used as the training set. The process is also repeated until all 4 groups have been set in turn as a testing group. We show the ROC analysis and A_z values of four sets in Fig. 12.

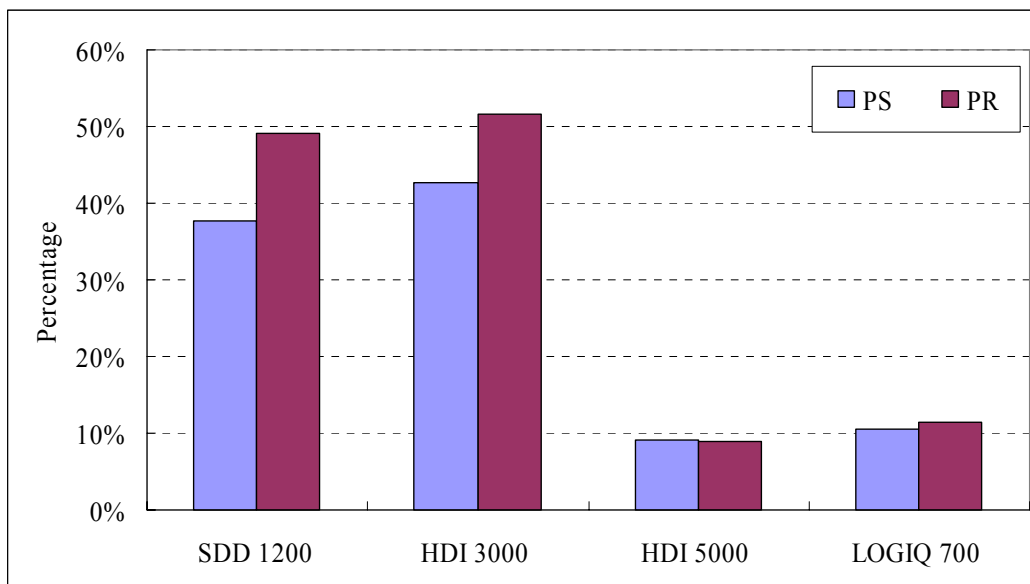


Fig. 11. The bar graph of the percentage of images of specific ultrasonic system in the entire database (denotes PS) and the Percentage of retrieved and query images from the identical ultrasonic system (denotes PR)

According to the Fig.12, the proposed method achieves reasonably high performances for all four sets of different ultrasonic systems in term of A_z value. The average diagnostic time for a breast US image was smaller than 5 milliseconds. The simulations were made on a single CPU Intel Pentium-4® 2.4 GHz personal computer with Microsoft Windows XP® operating system.

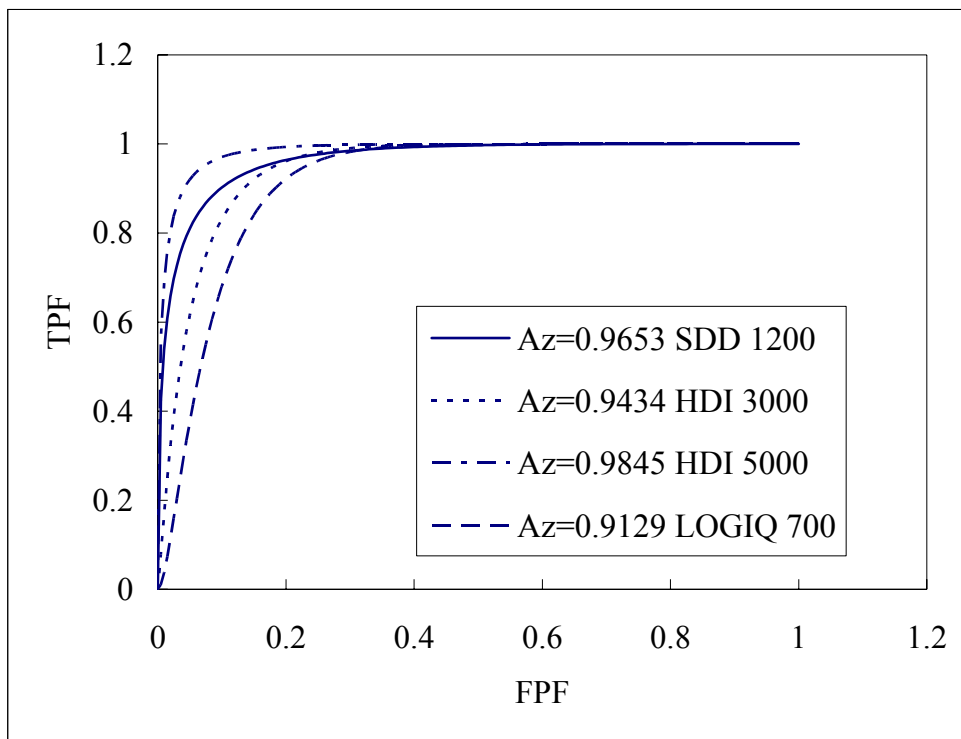


Fig. 12. The ROC analysis of all four sets of different ultrasonic systems

CHAPTER5

CONCLUSION AND DISCUSSION

This thesis proposes two efficient CAD systems to differentiate between benign and malignant tumors. Comparing to the previous works, the first CAD system using SVM is more efficiently than that using MLP. The MLP classifier ($A_Z = 0.956$) proposed by Chen et al. [9] is replaced by the SVM to obtain a better result ($A_Z = 0.970$). From the highly satisfactory specificity and sensitivity of results, the proposed system is expected to be a helpful tool for classifying benign and malignant tumors in sonograms, and can provide a second reading to help reduce misdiagnosis. With the growth of the database, more and more cases will be collected and used as training set. The MLP system suffered from the time consuming and initial condition dependent problems. The SVM system can improve those defects of MLP. Experimental results demonstrate the feasibility and excellent performance of the proposed system in sonogram classification.

The second CAD system uses the image retrieval technique to diagnose breast cancer on differential ultrasonic systems. Four different ultrasonic systems are used in the current medical diagnosis with the rapid development of sonogram technologies. Resolution, contrast, and so on are the main concerns for designing a CAD system for different ultrasonic systems. How to transform the needed information for diagnosis

between different systems becomes the most important issue. The users care about whether a designed system was suitable to another US machine without any change or through the adjustment of the parameters. Those change and adjustment of the parameters depend by using intelligent selection algorithms according to the different US machines. Previously Chen et al. [30] proposed the novel diagnosis system for different ultrasonic systems which inter-pixel correlation in the US images was used to differentiate benign and malignant tumors. The information needed for diagnosis between two different systems is achieved through the proposed adjustment technique [30]. However, this still need to have adjustment schemes for different ultrasonic systems. The further efforts must be made to transform the information among them if the more different ultrasonic systems existed. The image retrieval technique utilizes the projected principal vector to query the texture similar US images from database. The methods avoided the perplexity training procedure. Furthermore, historical cases can be directly added into the reference database without retraining. With the growth of the database, the new cases can be collected and used as references easily. The bar graph in Fig. 12 indicates that the diagnostic performance of the proposed system was not only supported by the query and retrieved images that were acquired from the same ultrasonic system. Moreover, Fig. 12 shows that a correct diagnosis may be made by referring the retrieved images from the different ultrasonic scanners.

REFERENCES

- [1] "Breast Cancer Facts and 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] P. Skaane and K. Engedal, "Analysis of sonographic features in the differentiation of fibroadenoma and invasive ductal carcinoma," *AJR Am. J. Roentgenol.*, vol. 170, no. 1, pp. 109-114, Jan. 1998.
- [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, 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.
- [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] 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.

- [8] 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.
- [9] S. Haykin, *Neural Networks: a comprehensive foundation*, 2 ed. NJ: Prentice Hall, 1999.
- [10] 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.
- [11] V.N. Gudivada and V.V. Raghavan, "Content-Based Image Retrieval-Systems," *Computer*, vol. 28, no. 9, pp. 18-22, Sept. 1995.
- [12] 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.
- [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] 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.
- [15] 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.
- [16] U. Sinha and H. Kangaroo, "Principal component analysis for content-based image retrieval," *Radiographics*, vol. 22, no. 5, pp. 1271-1289, Sept. 2002.

- [17] 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.
- [18] V. Vapnik, *Statistical Learning Theory* New York: John Wiley & Sons, 1998.
- [19] N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* United Kingdom: Cambridge University Press, 2000.
- [20] I. El Naqa, Y.Y. Yang, M.N. Wernick, N.P. Galatsanos, and R.M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1552-1563, Dec. 2002.
- [21] M.H. Yang, D. Roth, and N. Ahuja, "A tale of two classifiers: SNoW vs. SVM in visual recognition," *Computer Vision - Eccv 2002, Pt Iv*, vol. 2353 pp. 685-699, 2002.
- [22] K.I. Kim, K. Jung, S.H. Park, and H.J. Kim, "Support vector machines for texture classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1542-1550, Nov. 2002.
- [23] Y.F. Sun, X.D. Fan, and Y.D. Li, "Identifying splicing sites in eukaryotic RNA: support vector machine approach," *Computers in Biology and Medicine*, vol. 33, no. 1, pp. 17-29, Jan. 2003.
- [24] M.H. Song, C.M. Breneman, J.B. Bi, N. Sukumar, K.P. Bennett, S. Cramer et al., "Prediction of protein retention times in anion-exchange chromatography systems using support vector regression," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1347-1357, Nov. 2002.
- [25] Q. Song, W.J. Hu, and W.F. Xie, "Robust support vector machine with bullet hole image classification," *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, vol. 32, no. 4, pp. 440-448, Nov. 2002.

- [26] I.T. Jolliffe, *Principal Component Analysis* New York: Springer-Verlag, 1986.
- [27] 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.
- [28] R.C. Gonzalez and R.E. Woods, "Image Enhancement in the Spatial Domain," in *Digital image processing 2* ed. Massachusetts: Addison Wesley, 2002, pp. 75-146.
- [29] R.C. Gonzalez and R.E. Woods, "Image Compression," in *Digital image processing 2* ed. Massachusetts: Addison Wesley, 2002, pp. 409-518.
- [30] W.J. Kuo, R.F. Chang, W.K. Moon, C.C. Lee, and D.R. Chen, "Computer-aided diagnosis of breast tumors with different US systems," *Acad. Radiol.*, vol. 9, no. 7, pp. 793-799, July 2002.