AD_____

Award Number: DAMD17-02-1-0214

TITLE: Digital Mammography: Development of an Advanced Computer-Aided Diagnosis System for Breast Cancer Detection

PRINCIPAL INVESTIGATOR: Heang-Ping Chan, Ph.D.

CONTRACTING ORGANIZATION: The University of Michigan
Ann Arbor, Michigan 48109-1274

REPORT DATE: May 2005

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2005	Annual (1 May 2004 - 30 Apr 2005)

4. TITLE AND SUBTITLE
Digital Mammography: Development of an Advanced Computer-Aided Diagnosis System for Breast Cancer Detection

5. FUNDING NUMBERS
DAMD17-02-1-0214

6. AUTHOR(S)
Heang-Ping Chan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Michigan
Ann Arbor, Michigan 48109-1274

E-Mail: chanhp@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The goal of the project is to develop computer-aided diagnosis (CAD) methods and systems for mammography using advanced computer vision techniques and image information fusion from multiple mammograms to improve lesion detection and characterization. When fully developed, the CAD system can assist radiologists in mammographic interpretation.

During this project year, we have performed the following tasks: (1) collected databases of digital mammograms (DMs) and digitized film mammograms (DFMs) for development of the CAD systems, (2) developed computer vision techniques and a prototype CAD system for detection of microcalcifications on DMs, (3) developed computer vision techniques and a prototype CAD system for detection of masses on DFMs and DMs, (4) explored the feasibility of improving mass detection by CAD on digital breast tomosynthesis mammograms, (5) developed automated pectoral muscle detection method for preprocessing of MLO view mammograms for multiple-image analysis, (6) developed a prototype CAD method for classification of malignant and benign masses by fusion of information from mammograms and ultrasound images and investigated the effects of the multi-modality CAD system on radiologists' performance.

In summary, we have investigated a number of areas in CAD of mammographic lesions and evaluated the new techniques for both DMs and DFMs. We have made progress in the six tasks proposed in the project. We have found that our new computer-vision techniques and two-view information fusion approach can improve the performance of the CAD systems. We will continue the development of the CAD systems for DMs and DFMs in the coming years.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Breast cancer, digital mammography, computer-aided diagnosis, breast cancer diagnosis	68
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

(3) Table of Contents

(4) Introduction

Computer-aided diagnosis (CAD) has been shown to be useful as a second opinion to radiologists for breast cancer detection on mammograms. All current CAD systems have been developed for digitized screen-film mammograms (DFM). With the recent advent of full field digital mammography (FFDM) systems, it is important to develop CAD systems specifically designed for direct digital mammograms (DMs) in order to fully exploit the advantages of FFDM. Although many computer vision techniques developed for digitized films may be used for DMs, proper adaptation and extensive training of the current algorithms for the new type of images will be required. More importantly, new techniques still need to be developed to further improve the current algorithms for DFMs as well as for adapting to FFDM.

The goal of the proposed research is to develop a CAD system for breast cancer diagnosis using advanced computer vision techniques. The proposed CAD system will assist radiologists with detection and classification of breast lesions. Previous CAD methods for lesion detection and characterization are generally based on image features extracted from a single view. Our proposed approach is based on two steps: the first step uses single view detection to identify lesion candidates on individual mammograms, the second step is to fuse image information from multiple views to reduce false positives and thus to improve the overall accuracy. Although the main goal of this project is to develop a CAD system for DMs, we plan to extend the CAD development to DFMs for the following reasons: (1) digital mammography only became available in the last few years, multiple-view film mammograms with breast lesions are more commonly available in existing patient files, and (2) screen-film mammography will still be the main modality for breast cancer screening in the near future. Therefore, we will first develop the multiple-view correlation techniques for the CAD system of the DFMs. These new techniques will then be adapted to the CAD system for DMs. We believe that this approach is more efficient and we will obtain a CAD system for DMs as well as improve the CAD system for DFMs.

The following specific aims will be addressed: (1) Collection of databases of both DMs and DFMs and design of a database management system. (2) Improvement of single-view computer vision techniques for mass detection and classification in DFMs. (3) Improvement of single-view computer vision techniques for microcalcification detection and classification in DFMs. (4) Development of methods for correlation of image information from two-view DFMs. (5) Comparison of the detection and classification accuracy of the multiple-view fusion CAD system with the performance of the single-view CAD system by receiver operating characteristic (ROC) and FROC analyses. (6) Adaptation of the computer vision techniques to the CAD system for DMs. (7) Adaptation of the multiple-view fusion methods to the CAD system for DMs.

We will develop novel regional registration methods for identifying corresponding lesions on craniocaudal (CC) and mediolateral oblique (MLO) views. The multiple image information will be fused with specially designed correspondence classifiers or fuzzy classification to reduce false positives and to improve lesion detection sensitivity. Multiple-view features of a lesion will be merged using neural networks or other classifiers for classification of malignant and benign lesions. In addition, new computer vision techniques will be developed in each of the four areas to improve the current methods. The techniques will be first developed for DFMs. The algorithms for DFMs will then be adapted to DMs, taking into account the differences in the imaging characteristics between DMs and DFMs. Databases of DFMs and DMs will be collected from our patient population with IRB approved protocol and extensive training and independent testing of the new CAD system will be performed. The test performance of the multiple-image correlation CAD algorithms for detection and characterization of lesions on DFMs will be compared with the one-

view approach on DFMs as well as the performances of CAD systems for DMs using ROC methodology.

DM or DFM not only has the potential to detect breast cancer in an early stage, it will also facilitate consultation via teleradiology in remote or rural regions where expert mammographers may not be readily available. An effective CAD system will be particularly useful for providing an additional on-site or remote second opinion. This will be highly relevant to women in the military, especially when they are stationed in remote areas. DM in combination with CAD will fully utilize the potential of mammography to improve the health care of women both in the military and in the general population.

(5) Body

This is the third year annual report of our project. In the project period (5/1/04-4/30/05), we have extended our investigations to both the CAD systems for DMs and DFMs, and performed a number of studies to develop the CAD system for breast cancer diagnosis. A summary of some of the important accomplishments follows.

(A) Collection of databases of digital mammograms and digitized film mammograms

We continue to collect the database of digital mammograms (DMs) with mammographic masses or clustered microcalcifications for the development of our computer-aided diagnosis (CAD) algorithms. We have collected about 220 cases containing more than 800 mammograms. The patients were diagnosed with lesions in their mammograms during their normal clinical care, either by routine screening or by referral to our breast imaging clinic for evaluation. Most of the cases contained both DMs and screen-film mammograms.

As described in our previous reports, the digital mammograms are acquired with a GE Senographe 2000D full field digital mammography (FFDM) system. The pixel size of the system is 100 μm X 100 μm. The gray level resolution of the system is 14 bits for the raw images and 12 bits for the processed images. After acquisition, the digital image files are transmitted to the Siemens Archive which is the PACS system used in our department for storage of all clinical digital images. With Institutional Review Board (IRB) approval, we download the DMs from the Siemens Archive to our laboratory and digitize the film mammograms from the same patient. The film mammograms are digitized with a Lumiscan 85 laser scanner at a pixel size of 50 μm X 50 μm and a 12 bit gray level. We have developed a database management program based on Microsoft Access to process the images downloaded to our system. For each mammogram file, all patient identifiers are first removed from the image header. The patient name is replaced with a code number. The image is then named by the code number, the view (craniocaudal, mediolateral oblique, or mediolateral), and the exam year. A record is generated in the database file for each image. The record keeps the code number, the lesion type, the view, and the exam date information for each case. If the pathology of the case is available, the malignant or benign information of the lesion is also entered. Each case in the database will be read by an experienced MQSA radiologist to mark the lesion location. For microcalcification cases, the radiologist measures the diameter of the cluster, and provides description of its distribution, morphology, and visibility of the microcalcifications. For mass cases, the radiologist measures the diameter of the mass, and provides description of its margin, shape, spiculated or non-spiculated, the visibility, and the density of the mass relative to that of the parenchyma. For all cases, the

radiologist also provides BI-RADS description of the breast density and estimates the likelihood of malignancy of the lesion. These descriptions are entered into the database for each case as a reference for future analysis.

(B) CAD system for microcalcification detection on digital mammograms

We are developing a computer-aided detection (CAD) system to detect microcalcification clusters automatically on DMs. In this study, we investigated the performance of a nonlinear multiscale Laplacian pyramid enhancement method in comparison with a band-pass box-rim filter at the image enhancement stage and the use of a new error metric to improve the efficiency and robustness of the training of a convolution neural network (CNN) at the FP reduction stage of our CAD system.

Methods:

Our CAD system includes five stages: preprocessing, image enhancement, segmentation of microcalcification candidates, false positive (FP) reduction based on a convolution neural network (CNN), and regional clustering. In this study, the microcalcification detection system previously developed for DFMs was adapted to FFDMs by retraining. To develop a CAD system which is less dependent on the FFDM manufacturer's proprietary preprocessing methods, we used the raw FFDM images as input to our CAD system. The Laplacian pyramid enhancement method was described in detail in last year's report for our mass detection CAD system for DMs. A data set of 96 two-view mammograms containing 192 FFDMs acquired with a GE Senographe 2000D system. An MQSA radiologist identified the biopsy-proven cluster in each case. The 96 cases were separated into two independent subsets for cross validation training and testing. CNN training and validation were performed within one subset, and the performance of the trained system was evaluated on the independent test subset and quantified by free-response receiver operating characteristic (FROC) analysis.

The FROC curve shows the detection sensitivity (true positive fraction -TPF) as a function of FPs per image and thus shows the trade-off between sensitivity and specificity of a detection algorithm. FROC curves were presented on a per-image and a per-case basis. For image-based FROC analysis, the cluster on each mammogram was considered an independent true cluster. For case-based FROC analysis, the same cluster imaged on the two-view mammograms was considered to be one true cluster and the detection of either or both on the two views was considered to be a TP. The test FROC curve was obtained from averaging the test FROC curves from the two subsets.

Results:

For this data set, Laplacian pyramid multiscale enhancement did not improve the performance of the microcalcification detection system in comparison with our box-rim filter previously optimized for digitized screen-film mammograms. With the new error metric, the training of CNN could be accelerated and the classification performance in validation was improved from an A_z value of 0.94 to 0.97 on average. The CNN in combination with rule-based classifiers could reduce FPs with a small tradeoff in sensitivity. The test FROC curves are shown in Fig. 1. Our current CAD system for DMs can achieve an image-based sensitivity of 86% and 88%, respectively, at 0.49 and 0.85 FP marks/image. For case-based performance evaluation, a sensitivity of 90%, and 94% can be achieved at the same FP marker rates, respectively.

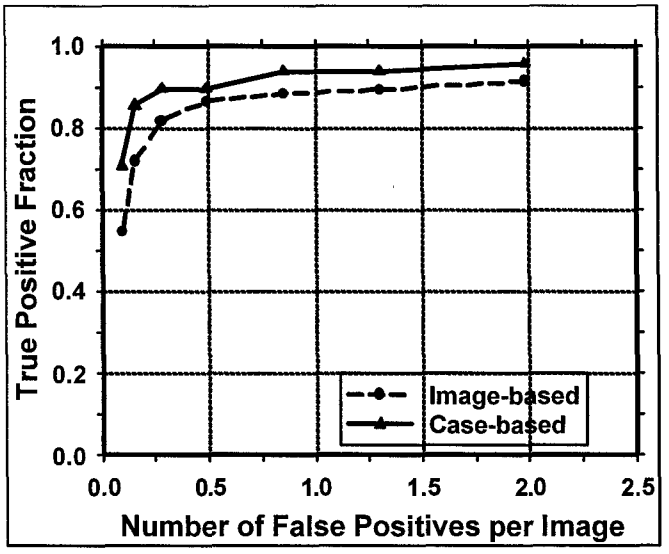


Fig. 1. Image-based and case-based FROC curve of the current microcalcification detection system.

Conclusion:

The performance of the CAD system for microcalcification detection on DMs has achieved a reasonable level in this data set. The DMs has relatively low noise and good signal-to-noise ratio so that the image enhancement with Laplacian pyramid multiscale method does not perform better than a simple band-pass filter. Further study is underway to collect a larger data set and to improve the performance of the system.

(C) CAD system for mass detection on digitized film mammograms

We have been investigating methods for improvement of the CAD system for detection of masses on DFMs as well as adapting the CAD system to mass detection on DMs. In this project year, we have developed a new dual CAD system approach which combines a regular CAD system with a new system trained with masses seen on retrospective review of prior mammograms to improve its performance for detecting subtle masses. The study is summarized below.

Methods:

Our regular CAD system for mass detection consisted of five steps: preprocessing, image enhancement, clustering-based region growing and local refinement, extraction of morphological and texture features, and rule-based and linear classification for FP reduction. Preprocessing included breast boundary detection and identification of the breast region. Image enhancement was achieved by a gradient field analysis method that located mass candidates based on the locations where strong gradient converged radially towards a point. Gradient field feature and morphological features were incorporated into the linear classifier in the FP reduction step. The simplex optimization procedure was used in the stepwise feature selection process to select the most effective features for classification of true masses and normal breast tissues. This single system approach has been discussed in last year's annual report.

In this study, two single CAD systems were optimized separately, one with current mammograms and the other with prior mammograms. A two-stage gradient field analysis was used to prescreen for mass candidates in both CAD systems. The suspicious structure in each identified region was extracted by clustering-based region growing. Morphological and spatial gray-level dependence texture features were extracted from the current mammograms. For detection of the subtle masses on the prior mammograms, histogram and run-length statistics features were extracted. Stepwise linear discriminant analysis (LDA) with simplex optimization was used to select the most useful features. Finally, rule-based and LDA classifiers were used to differentiate masses from normal tissues. When the dual CAD system was applied to a given mammogram, the detection information by the two CAD systems on the same mammogram were merged with a fusion scheme. The dual CAD system approach is illustrated in Fig. 2.

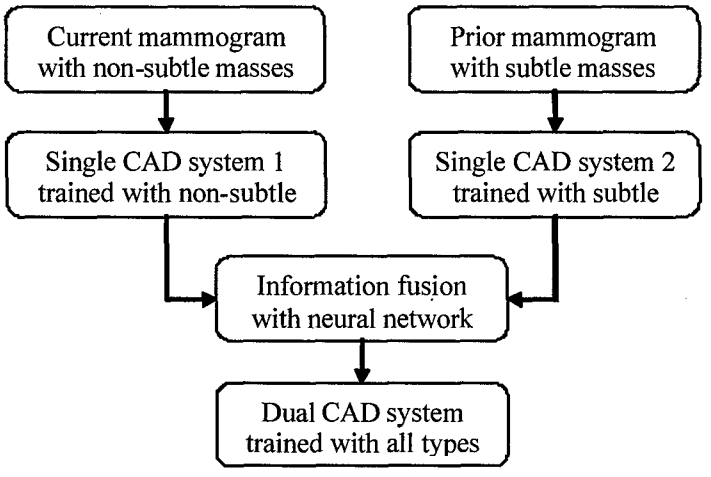


Fig. 2. Schematic of a dual system for mass detection in mammograms.

A data set 172 current mammograms containing biopsy-proven masses and 214 prior mammograms from 86 patients was used. The mammograms were digitized by a Lumiscan laser scanner with a pixel size of 100 μm X 100 μm and 12 bits per pixel. All of the current cases had two mammographic views: the craniocaudal (CC) view and the mediolateral oblique (MLO) view or the lateral view. We randomly separated the cases in our data set into two independent equal sized data sets, each with 43 cases. The training and testing were performed using the cross validation method. FROC curves were presented on a per-image and a per-case basis. The average test FROC curve was obtained by averaging the FP rates at the same sensitivity along the two corresponding test FROC curves from the 2-fold cross validation.

Results:

When the single CAD system trained on the average data set was applied to the test set, the A_z for FP classification was 0.81 and the FP rates were 2.1, 1.5 and 1.3 FPs/image at the case-based sensitivities of 95%, 90% and 85%, respectively. With the dual CAD system, the A_z was 0.85 and the FP rates were improved to 1.7, 1.2 and 0.8 FPs/image at the same case-based sensitivities. Fig. 3 shows the comparison of the test performance of the single and dual CAD

systems by using image-based and case-based average FROC curves. A comparison of the FP rates at several sensitivities for the single and dual systems is shown in Table 1.

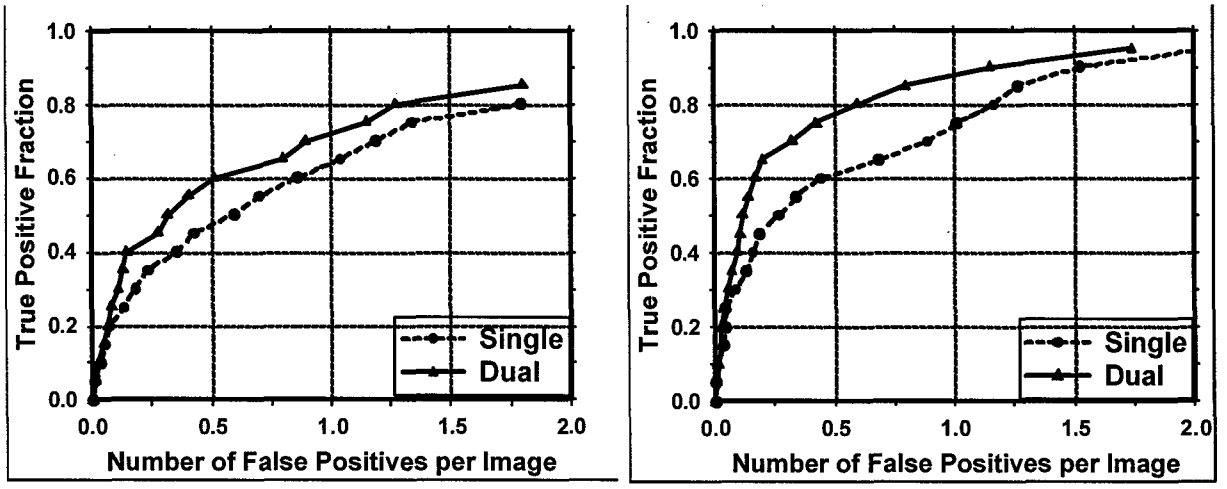


Fig. 3. Image-based (left) and case-based (right) average FROC curves obtained from averaging the corresponding FROC curves of the two test subsets. Single: detection by the single CAD system. Dual: detection by the dual CAD system.

Table 1. Comparison of detection accuracy for our single CAD system and our new dual CAD system for mass detection. TPF=true positive fraction, FP=false positive.

Image-based Scoring			Case-based Scoring		
TP	FP/image		TP	FP/image	
	Single	Dual		Single	Dual
85%		1.81	95%	2.08	1.74
80%	1.80	1.28	90%	1.53	1.15
75%	1.34	1.15	85%	1.27	0.80

Conclusion:

The dual CAD system could achieve a higher accuracy than the single CAD system. The dual system approach is a promising method for improvement of the accuracy of detecting subtle masses. Further study is underway to optimize the fusion scheme in our dual system.

(D) Computer-aided detection on digital breast tomosynthesis (DBT) mammograms

DBT is a new modality that holds the promise of improving breast cancer detection. Although it is not included in our original proposal, we believe that this new modality in

combination with CAD will be an exciting new direction for improving breast cancer detection and diagnosis. We thus performed a pilot study to investigate the feasibility of developing a CAD system for breast masses on digital breast tomosynthesis (DBT) mammograms.

Materials and Methods:

A schematic of the CAD system for DBT mammograms is shown in Fig. 4. For an input case of DBT slices, the computer detection system first screened the 3D volume for mass candidates by gradient field analysis. Each mass candidate was segmented from the surrounding structured background in the volume of interest. Morphological, gray level, and texture features were then extracted from the segmented structures. A feature classifier using linear discriminant analysis was designed to differentiate true masses from normal tissues.

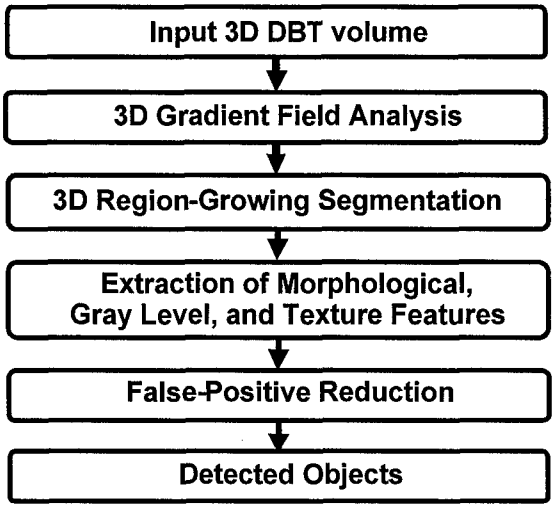


Fig. 4. Schematic of a CAD system for mass detection in DBT mammograms.

In this pilot study, we used 26 DBT cases which were acquired with a GE DBT prototype system at the Breast Imaging Research Lab of Massachusetts General Hospital (MGH). All cases were obtained with the approval of the Institutional Review Board at the MGH. The patients (age range: 41-77, mean =56, median=56) were recruited with written informed consent. The DBT system acquired 11 projection views (PVs) of the compressed breast over a 50-degree arc in the MLO view. DBT slices were reconstructed at 1-mm slice spacing from the PVs using an iterative maximum-likelihood algorithm. The cases included 23 masses and 3 areas of architectural distortion. The number of DBT slices per case ranged from 37 to 89 (mean=60.1). The CAD system was trained and tested using a leave-one-case-out resampling method. The detection accuracy was evaluated by FROC analysis.

Results:

The classification accuracy of the feature classifier for false-positive reduction reached an area under the receiver operating characteristic curve of 0.91 ± 0.03. The CAD system achieved a sensitivity of 85% at 2.2 false positives/case with this limited data set. The FROC curve obtained from leave-one-out testing is shown in Fig. 5.

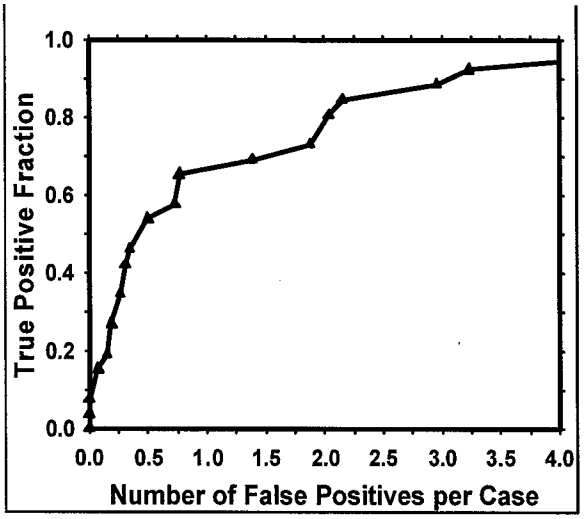


Fig. 5. Image-based and case-based FROC curve of the current microcalcification detection system.

Conclusion:

The preliminary results demonstrate the feasibility of our approach to the development of a CAD system for assisting radiologists in detecting masses on DBT mammograms.

(E) Computerized pectoral muscle identification on MLO-view mammograms for CAD applications

Automatic identification of the pectoral muscle on MLO view is an essential step for computerized analysis of mammograms. It can reduce the bias of mammographic density estimation, will enable region-specific processing in lesion detection programs, and also may be used as a reference in image registration algorithms. We are developing a computerized method for the identification of pectoral muscle on mammograms.

Methods:

A schematic of the automated pectoral muscle detection system is shown in Fig. 6. The upper portion of the pectoral edges was first detected to estimate the direction of the pectoral muscle boundary. A gradient-based directional (GD) filter was used to enhance the linear texture structures, and then a gradient-based texture analysis was designed to extract a texture orientation image that represented the dominant texture orientation at each pixel. The texture orientation image was enhanced by a second GD filter. An edge flow propagation method was developed to extract edges around the pectoral boundary using geometric features and anatomic constraints. The pectoral boundary was finally generated by a second-order curve fitting. We used 118 MLO view mammograms in this study. An experienced radiologist traced the pectoral muscle boundary on each image using a graphical user interface. They were used as gold standard for evaluation of the computer performance.

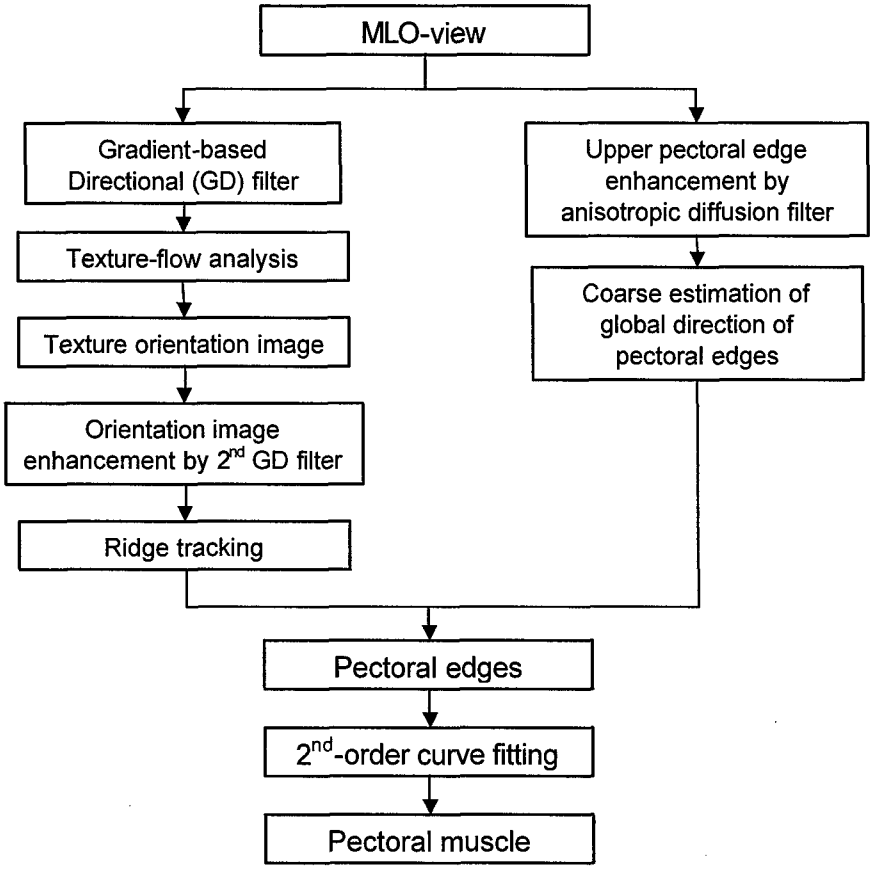


Fig. 6. Automated pectoral muscle detection scheme.

Results:

For each MLO view mammogram, the accuracy of pectoral boundary detection was evaluated by two performance metrics: the percentage of overlap, defined as the ratio of the overlap area between the computer detected pectoral muscle area and the gold standard relative to the gold standard, and the root-mean-square (RMS) distance obtained by calculating the shortest distance point by point between the computer-identified pectoral boundary and the manually marked pectoral boundary. For the data set of 118 MLO view mammograms, 99.2% (117/118) of the pectoral muscles could be identified. The average of the percent overlap area is 94.8% with a standard deviation of 20.9%. The average of the RMS distance is 4.3 mm with a standard deviation of 5.9 mm.

Conclusion:

The results indicate that the pectoral muscle on mammograms can be detected accurately by our automated method. The newly developed gradient-based directional filter and the dominant texture orientation estimation method can enhance the pectoral boundary regions. The edge flow propagation method can accurately extract pectoral edges to generate the pectoral boundary. We plan to incorporate this method into our CAD system for processing MLO mammograms and for multiple image analysis.

(F) Effect of a multi-modality classifier on radiologists' accuracy in characterizing breast masses on mammograms and ultrasound images

In clinical practice, it is known that ultrasound (US) images of breast masses are useful adjunct to mammograms for differentiation of malignant and benign masses. In our project, we proposed to develop a classifier for assisting radiologists in estimating the likelihood of malignancy of masses on mammograms. We have a prototype mass classification system for mammographic masses. In this preliminary study, we evaluated the effectiveness of merging the information from the two modalities, and assessed its effect on radiologists' accuracy in a receiver operating characteristic (ROC) study.

Methods:

A schematic of the CAD system that combines the information extracted from the mammograms and US images of a mass is shown in Fig. 7. The 3D US volumetric data were collected as cine-clips when the transducer was translated across the lesion. The masses were automatically segmented by the computer in each modality. US features were extracted based on the margin, shadowing, and shape characteristics. Mammographic features were extracted based on texture, morphological, and spiculation characteristics. We compared different techniques for combining the features or computer scores from the two modalities. The classifier was trained and tested with a leave-one-case-out method.

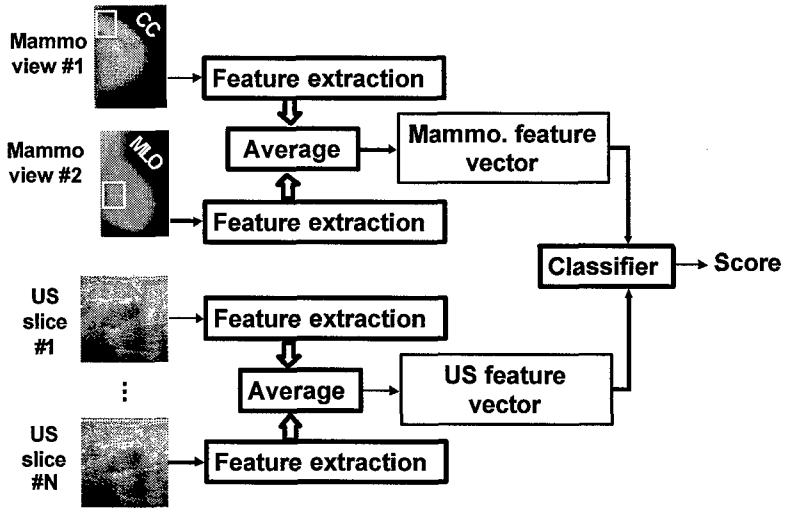


Fig. 7. A schematic representation of the classifier that combines the information extracted from mammograms and ultrasound images of breast masses.

After the classifier is developed, an observer performance study was performed to evaluate the effects of CAD on radiologists' classification of malignant and benign masses. Five MQSA radiologists participated as observers in the ROC study. First, the radiologist read the mammograms, and provided a BIRADS score and a malignancy rating for the mass. Second, the US images were displayed along with the mammograms, the radiologist provided a second malignancy rating and recommended one of three action categories: (i) 1-year follow-up; (ii) short-term follow-up; and (iii) biopsy. Third, the computer score was displayed, and the

radiologist provided a third malignancy rating and revised the recommended action. The classification accuracy was quantified using the area under ROC curve, A_z.

The data set consisted of images from 67 patients containing biopsy-proven masses (32 benign and 35 malignant). None of the masses were simple cysts. An experienced radiologist, using image data and patient reports, identified the region of interest (ROI) containing the lesion on both modalities. Each case contained 1 to 3 mammographic views and the 3D US volumetric data in digital images.

Results:

Table 2 summarizes the results of this study. For the computer classifier, the best information fusion methods achieved an A_z value of 0.84, 0.87, and 0.91, respectively, using mammogram alone, US alone, and both modalities. The results of the ROC study comparing radiologists' performances in classification of malignant and benign masses without and with CAD are also shown in Table 2. The radiologists had an average A_z of 0.87 reading mammograms alone. The average A_z was 0.93 when the mammograms were supplemented by US images. With CAD, the accuracy of every radiologist improved, and the average A_z increased to 0.95. The improvement was statistically significant ($p < 0.05$).

Table 2. Comparison of radiologists' performance in classification of malignant and benign masses without and with CAD. The computer classifier performance is also shown for reference.

Reading Condition	A_z	Sensitivity	Specificity
Computer: mammo	0.84		
Computer: US	0.87		
Computer: mammo+US	0.91		
Radiologist: mammo	0.87	0.94	0.33
Radiologist: mammo+US	0.93	0.98	0.27
Radiologist: mammo+US+CAD	0.95	0.99	0.29
Radiologist: mammo+US+CAD	0.95	0.98	0.39

Conclusion:

The computer classifier achieved a performance approaching that of the average of experienced breast radiologists. The radiologists are more accurate in characterizing breast masses when both mammograms and US images are available. Although the radiologists' achieved high accuracy in the multi-modality reading condition, a well-trained computer algorithm can still improve radiologists' performance. The improved specificity by 12% indicates that CAD has the potential to reduce unnecessary biopsies.

(6) Key Research Accomplishments

- Continue collection of a database of digital mammograms and digitized film mammograms for development of the CAD algorithms for both digital mammography and film mammography ------ (Task 1)

- Develop improved computer vision techniques and a prototype CAD system for detection of microcalcifications on digital mammograms and evaluate the system performance by FROC analysis ------- (Task 3(a), Task 6(a))

- Develop improved computer vision techniques and a prototype CAD system for detection of masses on digitized and digital mammograms and evaluate the system performance by FROC analysis ------- (Task 2(a), Task 6(a))

- Explore the feasibility of improving mass detection by CAD on digital breast tomosynthesis mammograms (DBT) and evaluate the prototype system by FROC analysis ------- (Task 2(a))

- Develop automated pectoral muscle detection method that will serve as a preprocessing step for analysis of MLO view mammograms and for development of multiple-image analysis in an advanced CAD system ------ (Task 2, Task 3, Task 4, Task 5)

- Develop a prototype computerized method for classification of malignant and benign masses by fusion of information from mammograms and ultrasound images. Investigate the effects of the multi-modality CAD system on radiologists' performance on classification of masses ------ (Task 2(b), Task 4(b), Task 6(b))

(7) Reportable Outcomes

As a result of the support by the PRMRP grant, we have conducted studies in CAD for mammography and published the results. The publications in this project year are listed in the following.

Peer-Reviewd Journal Articles:

1. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Improvement of radiologists' characterization of malignant and benign breast masses in serial mammograms by computer-aided diagnosis: An ROC study. Radiology 2004; 233: 255-265.

2. Zhou C, Chan HP, Paramagul C, Roubidoux MA, Sahiner B, Hadjiiski LM, Petrick N, Computer-aided diagnosis on mammograms using multiple image analysis: computerized nipple identification. Medical Physics 2004; 31: 2871-2882.

3. Chan HP, Goodsitt MM, Helvie MA, Hadjiiski LM, Lydick JT, Roubidoux MA, Bailey JE, Nees A, Blane CE, Sahiner B. ROC study of the effect of stereoscopic imaging on assessment of breast lesions. Medical Physics 2005; 32: 1001-1009.

Accepted for Publication:

1. Chan HP, Wei J, Sahiner B, Rafferty EA, Wu T, Roubidoux MA, Moore RH, Kopans DB, Hadjiiski LM, Helvie MA. Computer-aided detection system for breast masses on digital tomosynthesis mammograms – Preliminary experience. Radiology. (in press)

2. Hadjiiski LM, Sahiner B, Helvie MA, Chan HP, Roubidoux MA, Paramagul C, Blane C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Computer-aided diagnosis of breast cancer in serial mammograms. Radiology.

Non-Peer-Reviewd Conference Proceeding Articles:

1. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Petrick N, Helvie MA, Zhou C, Ge Z. Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis. Proc SPIE 5370; 2004: 992-998.

2. Hadjiiski LM, Helvie MA, Sahiner B, Chan HP, Roubidoux MA, Nees A, Petrick N, Blane C, Paramagul C, Bailey J, Patterson S, Klein K, Adler D, Foster M, Shen J. ROC Study of the Effects of Computer-Aided Interval Change Analysis on Radiologists' Characterization of Breast Masses in Two-View Serial Mammograms. Proc SPIE 5370; 2004: 51-58.

Conference Proceedings:

1. Chan HP, Sahiner B, Hadjiiski LM. Sample size and validation issues on the development of CAD systems. In: CARS 2004 - Proc. 18th International Congress and Exhibition. Computer-Assisted Radiology and Surgery. Chicago, IL, June 23-26, 2004. Ed. Lemke HU, Vannier MW, Inamura K, Farman AG, Doi K, Reiber JHC. International Congress Series 1268. pp. 872-877. (Elsevier, Amsterdam, Netherlands).

2. Chan HP, Wei J, Sahiner B, Rafferty EA, Wu T, Roubidoux MA, Moore RH, Kopans DB, Hadjiiski LM, Helvie MA. Computerized detection of masses on digital tomosynthesis mammograms – A preliminary study. In: Digital Mammography IWDM 2004: 7th International Workshop on Digital Mammography. Ed. Pisano E. (in press).

3. Zhou C, Hadjiiski LM, Paramagul C, Sahiner B, Chan HP, Wei J. Computerized pectoral muscle identification on MLO-view mammograms for CAD applications. Proc SPIE 5747; 2005: (in press).

4. Hadjiiski LM, Chan HP, Sahiner B, Helvie MA, Roubidoux MA. Effects of the continuous and discrete confidence rating scales in ROC observer studies. Proc SPIE 5749; 2005: (in press).

5. Ge J, Wei J, Hadjiiski LM, Sahiner B, Chan HP, Helvie MA, Zhou C, Ge Z. Computer aided detection of microcalcification clusters on full-field digital mammograms: multiscale pyramid enhancement and false positive reduction using an artificial neural network. Proc SPIE 5747; 2005: (in press).

6. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Helvie MA, Roubidoux MA, Petrick N, Zhou C, Ge J. Computer aided detection of breast masses on mammograms: performance improvement using a dual system. Proc SPIE 5747; 2005: (in press).

7. Sahiner B, Hadjiiski LM, Chan HP, Zhou C, Wei J. Comparison of decision tree classifiers with neural network and linear discriminant analysis classifiers for computer-aided diagnosis: A Monte Carlo simulation study. Proc SPIE 5747; 2005: (in press).

Scientific/Education Exhibits

1. Sahiner B, Hadjiiski LM, Chan HP, Nees AV, Bailey JE, Blane CE, et al. Development of a multi-modality computer classifier for characterization of breast masses on mammograms and volumetric ultrasound images. Submitted for Education Exhibit at the 90th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 28-December 3, 2004.

Conference Abstracts and Presentations:

1. Chan HP, Wei J, Sahiner B, Rafferty EA, Wu T, Roubidoux MA, Moore RH, Kopans DB, Hadjiiski LM, Helvie MA. Computerized detection of masses on digital tomosynthesis mammograms – A preliminary study. Presented at the 7th International Workshop on Digital Mammography. IWDM-2004. Durham, North Carolina. June 18-21, 2004.

2. Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, LeCarpentier GL. Fusion of mammographic and sonographic computer-extracted features for improved characterization of breast masses. Presented at the 7th International Workshop on Digital Mammography. IWDM-2004. Durham, North Carolina. June 18-21, 2004.

3. Zhou C, Chan HP, Wei J, Helvie MA, Roubidoux MA, Paramagul C, Nees A, Hadjiiski LM, Sahiner B. Performance evaluation of an automated breast density estimation system for digital mammograms and digitized film mammograms. Presented at the 7th International Workshop on Digital Mammography. IWDM-2004. Durham, North Carolina. June 18-21, 2004.

4. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA. Automated interval change analysis of masses in serial mammograms - evaluation of an adaptive similarity measure for mass matching. Presented at the 7th International Workshop on Digital Mammography. IWDM-2004. Durham, North Carolina. June 18-21, 2004.

5. Chan HP, Wei J, Sahiner B, Rafferty EA, Wu T, Ge J, Roubidoux MA, Moore RH, Kopans DB, Hadjiiski LM, Helvie MA. Computer-aided detection on digital breast tomosynthesis (DBT) mammograms – Comparison of two approaches. Presented at the 90th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 28-December 3, 2004. RSNA Program 2004; 447.

6. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Helvie MA, Roubidoux MA. A dual computer-aided detection (CAD) system for improvement of mass detection on mammograms. Presented at the 90th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 28-December 3, 2004. RSNA Program 2004; 491.

7. Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul CP, Helvie MA, et al. The effect of a multi-modality computer classifier on radiologists' accuracy in characterizing breast masses on mammograms and ultrasound images: An ROC study. Presented at the 90th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 28-December 3, 2004. RSNA Program 2004; 447.

8. Hadjiiski LM, Chan HP, Sahiner B, Helvie MA; Roubidoux MA, Zhou C. Interval Change Analysis based on Computerized Regional Registration of Corresponding Microcalcification Clusters on Temporal Pairs of Mammograms. Presented at the 90th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 28-December 3, 2004. RSNA Program 2004; 491.

9. Zhou C, Hadjiiski LM, Paramagul C, Sahiner B, Chan HP, Wei J. Computerized pectoral muscle identification on MLO-view mammograms for CAD applications. Poster presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-17, 2005.

10. Hadjiiski LM, Chan HP, Sahiner B, Helvie MA, Roubidoux MA. Effects of the continuous and discrete confidence rating scales in ROC observer studies. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-17, 2005.

11. Ge J, Wei J, Hadjiiski LM, Sahiner B, Chan HP, Helvie MA, Zhou C, Ge Z. Computer aided detection of microcalcification clusters on full-field digital mammograms: multiscale pyramid enhancement and false positive reduction using an artificial neural network. Poster presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-17, 2005.

12. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Helvie MA, Roubidoux MA, Petrick N, Zhou C, Ge J. Computer aided detection of breast masses on mammograms: performance improvement using a dual system. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-17, 2005.

13. Chan HP, Hadjiiski LM, Ge J, Sahiner B, Helvie MA. Computer-aided diagnosis: computerized classification of malignant and benign microcalcifications on full field digital mammograms. Submitted for presentation at the 47th Annual Meeting of the American Association of Physicists in Medicine, Seattle, WA. July 24-28, 2005.

(8) Conclusions

Under the support of this grant, we have investigated various computer-aided detection and diagnosis (CAD) methods for analysis of lesions on mammograms. We continue to collect a database of digitized film mammograms (DFMs) and a database of full field digital mammograms (DMs) that contain mammographic lesions from our breast imaging division in the Department of Radiology. The digital images include the manufacturer's processed images and unprocessed (raw) images. All collected cases are entered into our database management program that stores the coded case information to facilitate archiving and retrieval of the cases.

As discussed in the annual report last year, we continue to develop computer-vision techniques using DFMs in parallel with DMs. These techniques should be readily transferable between DFMs and DMs with minor modifications. A main advantage of using DFMs for the development of CAD techniques is that many patients will have multiple exams on DFMs. The serial mammograms allow us to develop advanced techniques using subtle lesions that may have been overlooked in prior years or using interval change information for lesion detection or diagnosis. In this project year, we have developed a dual CAD system for mass detection that includes two single CAD systems, one trained with current mammograms and the other trained with prior mammograms. The fusion of information from the two systems was found to improve the accuracy of detecting subtle masses.

A new imaging modality using full field digital mammography system - digital breast tomosynthesis (DBT) mammography - is on the horizon. The 3D information available with the tomosynthesis mammograms holds the promise of improving breast cancer detectability, especially in dense breasts. We were able to obtain a small set of DBT mammograms through the collaboration with the research group at the Massachusetts General Hospital. We performed a pilot study of developing 3D computer-vision techniques for mass detection on DBT mammograms. We found that our techniques could achieve high accuracy in this small data set. This study indicates that CAD for DBT mammography is a promising approach to improving breast cancer detection.

We are developing computer-vision methods for classification of malignant and benign masses. In clinical practice, radiologists generally use ultrasound scans as an adjunct to assess breast masses. We studied the effects of combining mammographic and ultrasound image information on computerized classification and found that the multi-modality computer classifier out-performed the classifiers using information from mammographic images or ultrasound images alone. In an observer performance study, the radiologists' accuracy in classification of malignant and benign masses improved significantly when they read with CAD using the multi-modality classifier. The increased specificity indicates that CAD may be useful for reducing unnecessary

biopsies. Further study is underway to train the classifiers with a larger data set and to investigate their robustness in independent test cases.

The development of the CAD system for detection of microcalcifications is underway. In this project year, we performed studies to improve the computer vision techniques at the different stages of the system. We found that Laplacian pyramid multiscale enhancement did not improve the signal-to-noise ratio (SNR) of microcalcifications on FFDMs, while a trained convolution neural network could reduce the false positives significantly in the CAD system. Although the performance of the CAD system has reached over 90% sensitivity at less than 1 FPs per image in the current data set, further study is underway to train the system with a larger data set and to evaluate its generalizability to unknown cases.

Our advanced CAD systems will utilize multiple image information fusion to improve their performances. A number of preprocessing methods are being developed to register the lesions on multiple images of the same view and on multiple views. We previously developed an automated breast boundary detection method and an automated nipple detection method. In this project year, we investigated an automated method for detection of the pectoral muscle on MLO views. These image processing methods will be incorporated into our lesion registration algorithms for preparation of multiple image analysis.

In summary, we have investigated a number of areas in computer-aided detection and computer-aided diagnosis of mammographic lesions. We have made progress in the six tasks proposed in the project. This lays the strong foundation for us to continue the development of the CAD systems for digital mammograms and digitized film mammograms in the coming years.

(9) References

None.

(10) Appendix

Copies of the following publications are enclosed with this report.

Peer-Reviewd Journal Article:

1. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Improvement of radiologists' characterization of malignant and benign breast masses in serial mammograms by computer-aided diagnosis: An ROC study. Radiology 2004; 233: 255-265.

2. Zhou C, Chan HP, Paramagul C, Roubidoux MA, Sahiner B, Hadjiiski LM, Petrick N, Computer-aided diagnosis on mammograms using multiple image analysis: computerized nipple identification. Medical Physics 2004; 31: 2871-2882.

3. Chan HP, Goodsitt MM, Helvie MA, Hadjiiski LM, Lydick JT, Roubidoux MA, Bailey JE, Nees A, Blane CE, Sahiner B. ROC study of the effect of stereoscopic imaging on assessment of breast lesions. Medical Physics 2005; 32: 1001-1009.

Conference Proceedings:

1. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Petrick N, Helvie MA, Zhou C, Ge Z. Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis. Proc SPIE 5370; 2004: 992-998.

2. Hadjiiski LM, Helvie MA, Sahiner B, Chan HP, Roubidoux MA, Nees A, Petrick N, Blane C, Paramagul C, Bailey J, Patterson S, Klein K, Adler D, Foster M, Shen J. ROC Study of the Effects of Computer-Aided Interval Change Analysis on Radiologists' Characterization of Breast Masses in Two-View Serial Mammograms. Proc SPIE 5370; 2004: 51-58.

Radiology

Lubomir Hadjiiski, PhD
Heang-Ping Chan, PhD
Berkman Sahiner, PhD
Mark A. Helvie, MD
Marilyn A. Roubidoux, MD
Caroline Blane, MD
Chintana Paramagul, MD
Nicholas Petrick, PhD
Janet Bailey, MD
Katherine Klein, MD
Michelle Foster, MD
Stephanie Patterson, MD
Dorit Adler, MD
Alexis Nees, MD
Joseph Shen, MD

Index terms:
Breast neoplasms, diagnosis, 00.31, 00.32
Computers, diagnostic aid
Diagnostic radiology, observer performance

Published online before print
10.1148/radiol.2331030432
Radiology 2004; 233:255–265

Abbreviations:
A_z = area under ROC curve
$_{0.90}A'_z$ = partial A_z index
BI-RADS = Breast Imaging Reporting and Data System
CAD = computer-aided diagnosis
ROC = receiver operating characteristic
ROI = region of interest

[1] From the Department of Radiology, University of Michigan Medical Center, CGC B2102, 1500 E Medical Center Dr, Ann Arbor, MI 48109-0904 (L.H., H.P.C., B.S., M.A.H., M.A.R., C.B., C.P., J.B., K.K., M.F., S.P., D.A., A.N., J.S.); and Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, Md (N.P.). From the 2002 RSNA scientific assembly. Received March 17, 2003; revision requested June 13; final revision received January 9, 2004; accepted February 4. Supported by USAMRMC grants DAMD17-98-1-8211, DAMD17-02-1-0489, and DAMD17-02-1-0214. **Address correspondence to** L.H. (e-mail: *lhadjisk@umich.edu*).

Authors stated no financial relationship to disclose.

Author contributions:
Guarantor of integrity of entire study, L.H.; study concepts and design, L.H., H.P.C., B.S., M.A.H.; literature research, L.H., H.P.C., B.S.; experimental studies, M.A.R., C.B., C.P., J.B., K.K., M.F., S.P., M.A.H., D.A., A.N., J.S.; data acquisition, all authors; data analysis/interpretation, L.H., H.P.C., B.S., N.P.; statistical analysis, L.H.; manuscript preparation, definition of intellectual content, editing, revision/review, and final version approval, all authors

Improvement in Radiologists' Characterization of Malignant and Benign Breast Masses on Serial Mammograms with Computer-aided Diagnosis: An ROC Study[1]

PURPOSE: To evaluate the effects of computer-aided diagnosis (CAD) on radiologists' characterization of masses on serial mammograms.

MATERIALS AND METHODS: Two hundred fifty-three temporal image pairs (138 malignant and 115 benign) obtained from 96 patients who had masses on serial mammograms were evaluated. The temporal pairs were formed by matching masses of the same view from two different examinations. Eight radiologists and two breast imaging fellows assessed the temporal pairs with and without computer aid. The classification of accuracy was quantified by using the area under receiver operating characteristic curve (A_z). The statistical significance of the difference in A_z between the different reading conditions was estimated with the Dorfman-Berbaum-Metz method for analysis of multireader multicase data and with the Student paired *t* test for analysis of observer-specific paired data.

RESULTS: The average A_z for radiologists' estimates of the likelihood of malignancy was 0.79 without CAD and improved to 0.84 with CAD. The improvement was statistically significant ($P = .005$). The corresponding average partial area index was 0.25 without CAD and improved to 0.37 with CAD. The improvement was also statistically significant ($P = .005$). On the basis of Breast Imaging Reporting and Data System assessments, it was estimated that with CAD, each radiologist, on average, reduced 0.7% (0.8 of 115) of unnecessary biopsies and correctly recommended 5.7% (7.8 of 138) of additional biopsies.

CONCLUSION: CAD based on analysis of interval changes can significantly increase radiologists' accuracy in classification of masses and thereby may be useful in improving correct biopsy recommendations.

Breast cancer is one of the leading causes of death in the United States among women between 40 and 55 years of age (1). Mammography is currently the most sensitive method for detecting early breast cancer, and it is also the most practical for screening (2,3). Although general rules for differentiation between malignant and benign lesions exist, in clinical practice only 15%–30% of patients referred for biopsy are found to have a malignancy (4–6). Unnecessary biopsies increase health care costs and may cause patient anxiety and morbidity. It is therefore important to improve the accuracy of interpreting mammographic lesions, thereby improving the positive predictive values of mammography.

Radiologists routinely compare the current mammograms of a patient with those obtained in previous years, if available, for identifying interval changes, detecting abnormalities, and evaluating breast lesions. It is widely accepted that interval changes in mammographic features are very useful for detection of breast cancer (7,8). In a recent

study, Burnside et al (9) reported that in a diagnostic setting, comparison with the prior examination significantly ($P < .001$) increased the overall cancer detection rate.

A variety of computer-aided diagnosis (CAD) techniques have been developed to detect abnormalities and to distinguish malignant and benign lesions on mammograms. It has been shown that CAD systems could improve the radiologist's accuracy in both detection and characterization of breast lesions in a single mammographic examination.

Chan et al (10) performed an observer study to evaluate the effects of CAD, which was designed for characterization of malignant and benign masses on mammograms obtained from a single examination (11), on the radiologist's diagnostic accuracy. Two observer experiments were performed. In the first experiment, the radiologists evaluated a data set of masses on single-view mammograms. In the second experiment, they evaluated the masses on two-view mammograms. In both experiments, the radiologists' performance in terms of the area under receiver operating characteristic (ROC) curve (A_z) was significantly ($P = .022$ and .007, respectively) improved when reading with CAD was compared with reading without CAD.

Huo et al (12) developed a computer classifier for distinguishing between malignant and benign masses. Multiple views of the masses acquired in the same examination were used. An observer study with 12 radiologists was performed. The radiologists' performance in terms of the A_z was also significantly ($P = .001$) improved with computer aid.

Jiang et al (13) developed a computer classifier for classification of microcalcification clusters on multiple views of single-examination mammograms and also performed an observer study to evaluate its effectiveness. They found that with computer aid, the radiologists achieved a statistically significant ($P < .001$) improvement in the classification of microcalcifications. In addition, an increase in biopsy recommendations for malignant clusters, as well as a decrease in the recommendation of biopsy for benign lesions, was observed.

Authors of these previous studies of lesion classification with CAD used information from a single examination (11–17). When mammograms from multiple examinations are available, it can be expected that even higher accuracy may be achieved if the computer can utilize the information obtained from analysis of interval changes for the classification. We (18) have developed a classification scheme that combines prior and current information that is automatically extracted from masses on prior and current mammograms, respectively. We found that the classifier using the combined prior and current information performed significantly better ($P = .015$) in terms of the A_z than did the classifier using current information alone. Thus, the purpose of our study was to evaluate the effects of CAD on radiologists' characterization of masses on serial mammograms.

MATERIALS AND METHODS

Data Set

A set of 253 temporal pairs of mammograms containing biopsy-proved masses on the current mammograms was selected consecutively from our mammogram database, and the images were digitized. The mammograms were obtained from patients who had undergone biopsy of breast masses at our department. The data collection protocol had been approved by our institutional review board. Patient informed consent was waived for this retrospective study. The selection criterion was that the patient had undergone serial examinations in which a corresponding mass could be identified. The masses on both the current and prior mammograms encompassed a range of sizes and conspicuity that would be seen in clinical practice. We also tried to approximately balance the number of patients with malignant and benign masses. The data set consisted of 406 mammograms from 96 patients. The mammograms were digitized with a laser scanner (LUMISCAN 85; Lumisys, Los Altos, Calif) at a pixel resolution of 50 × 50 μm and 4096 gray levels. The digitizer was calibrated so that gray-level values were linearly proportional to the optical density in the range of 0–4, with a slope of 0.001 per pixel value. The digitizer output was linearly converted so that a large pixel value corresponded to a low optical density. The image matrix size was reduced by averaging every 2 × 2 adjacent pixels and was down-sampled by a factor of 2, resulting in images with a pixel size of 100 × 100 μm for further analysis.

There were 97 biopsy-proved masses (53 malignant and 44 benign) in 96 patients (age range, 37–86 years; mean, 59.6 years). One patient had a malignant mass in the left breast and a benign mass in the right breast. The 406 mammograms contained different mammographic views (193 craniocaudal, 177 mediolateral oblique, and 36 lateral) from multiple serial examinations of the masses, including those from the examination when the biopsy decision was made. By matching masses of the same view from two examinations, a total of 253 temporal pairs of images were formed, of which 138 had malignant and 115 had benign masses. In cases where there were only two examinations, a single pair was obtained for the given view. If there were three examinations, two or three temporal pairs were obtained. The distribution of the 253 temporal pairs among the 96 patients with 97 masses was as follows: 117 craniocaudal pairs originated from 87 masses, 115 mediolateral oblique pairs originated from 88 masses, and 21 lateral pairs originated from 17 masses. The same mass could have craniocaudal, mediolateral oblique, or lateral views. The prior mammogram was assessed as negative, benign, or probably benign in the prior year examination, and the majority remained so in retrospect. When a mass was not discretely visible on the prior mammogram, a Mammography Quality Standards Act–approved radiologist (M.A.H.), with 17 years of experience reading mammograms, defined the area where the mass would develop.

Since all 97 masses in this data set had undergone biopsy, the benign masses in this set could not be prospectively distinguished clinically from the malignant masses based on current mammographic criteria. The radiologists might have observed changes in or suspicious features of the benign masses that prompted them to recommend biopsy.

For the malignant masses, the average mass size was 8.0 mm on the prior and 11.5 mm on the current mammogram. The corresponding sizes were 9.9 and 11.5 mm, respectively, for the benign masses.

To simulate a more realistic clinical situation in which a radiologist also has to distinguish mass-mimicking fibroglandular tissue from true masses, 34 additional temporal pairs containing corresponding normal structures on the serial mammograms were also included. These normal structures were selected by an experienced radiologist (M.A.H.) and were deemed to be difficult to distinguish from masses without further diagnostic work-up. The main reason for inclusion of temporal pairs containing normal structures was to reduce potential bias the radiologists might have when they evaluated the cases in an ROC experi-

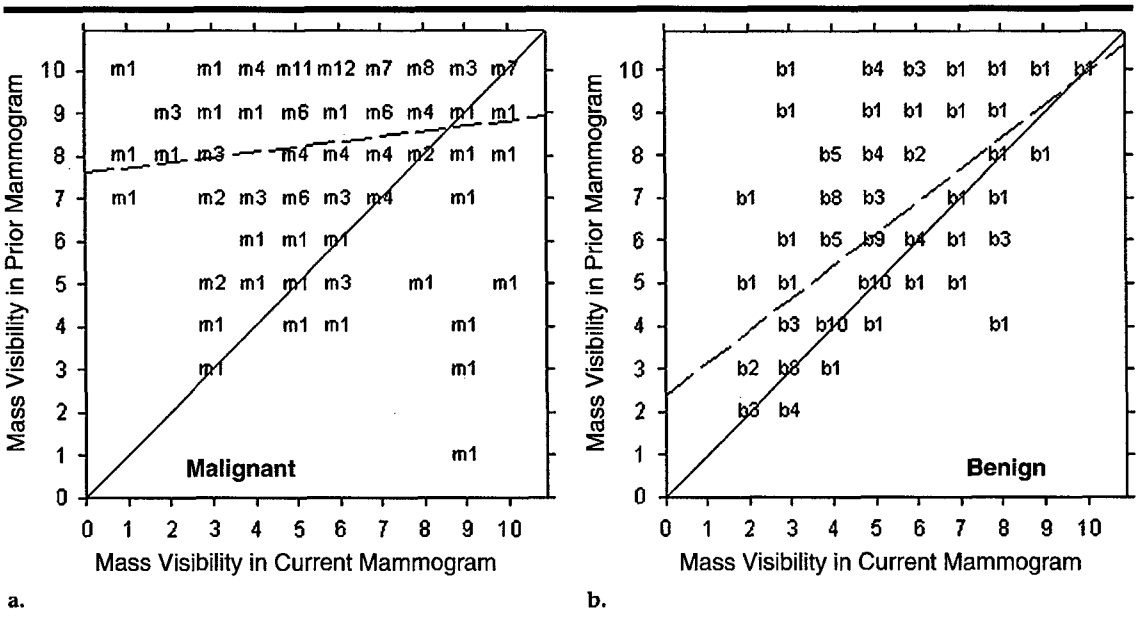


Figure 1. Graphs depict mass visibility on current and prior mammograms for **(a)** malignant and **(b)** benign temporal pairs of mammograms. Visibility was rated with a 10-point discrete scale (1, most obvious; 10, most subtle). Because many data points overlap, the number of points with the same rating are indicated by a number next to the symbol *m* or *b*. Diagonal line represents cases when the visibility ratings of current and prior masses are identical. The dashed linear regression line for the data is defined by **(a)** $y = 0.121x + 7.599$ and **(b)** $y = 0.755x + 2.367$. The correlation coefficient is 0.02 for malignant masses and 0.31 for benign masses.

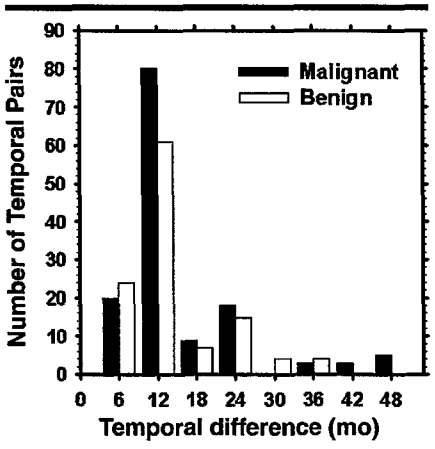


Figure 2. Histogram illustrates the temporal interval between current and prior mammograms for the 253 temporal pairs of mammograms in the data set.

ment. If the data set contained only malignant and benign masses (without normal pairs), the radiologists might be biased and give more optimistic scores. However, the 34 temporal pairs containing normal structures were excluded from the data analysis. In the analysis of results, it is more important to study the improvement in radiologists' performance when true masses are read. Therefore, all analyses were based on the 253 temporal pairs containing masses.

The radiologist also rated the visibility of the masses on the mammograms relative to those encountered in clinical practice by using a 10-point scale, with a score of 1 representing the most obvious and a score of 10 representing the most subtle masses. For the malignant and benign temporal pairs, the visibility of the masses on the current mammogram is plotted against that observed on the prior mammogram, as shown in Figure 1. Generally, the malignant masses were less visible on the prior than on the current mammogram, while the visibility of the benign masses was found to be more similar on the current and prior mammograms. The mean difference in the visibility ratings between prior and current mammograms for the malignant masses was 2.3 compared with 1.0 for the benign masses ($P < .001$ with an unpaired t test between the malignant and benign masses). The correlation coefficient was 0.02 for malignant masses (Fig 1a) and 0.31 for benign masses (Fig 1b). The temporal pairs had an interval of 6–48 months (Fig 2).

Computerized Classification of Temporal Masses

We have developed a classification technique that incorporates current and prior information to characterize the masses. The classification technique has been described in detail elsewhere (18). Figure 3 contains a flowchart of the method, which is summarized as follows: Initially, a region of interest (ROI) containing the mass was identified by a radiologist on both the current and prior mammograms. Automatic segmentation of the mass within each ROI was performed on the basis of a two-dimensional active contour model that was initialized with k-means clustering (19,20). Features related to texture, morphology, and spiculations were extracted from each mass (Appendix). A total of 35 features (20 based on run-length statistics, 12 morphological, and three spiculation) were extracted from each ROI. In addition, difference features were obtained by subtracting a prior feature from the corresponding current feature (Fig 3). Therefore, 35 difference features were derived from the 20 based on run-length statistics, 12 morphological features, and three spiculation features.

For the training and testing of the classifier, a "leave one case out" resampling scheme was used. To design a robust classifier, a subset of features was selected to reduce the dimensionality of the feature space. A stepwise feature selection was applied. For the 96 training subsets of temporal pairs used in this study, an average of seven features were selected for

the classification. The most frequently selected features included two difference features based on run-length statistics (gray-level nonuniformity and short-run emphasis), three current features based on run-length statistics (short-run emphasis, run-length nonuniformity, and long-run emphasis), one spiculation feature from the current image, and one spiculation feature from the prior image. The distribution of the classifier test score is presented in Figure 4. Small values correspond to benign ratings and large values correspond to malignant ratings. By using the ROC methods, the overall performance of the classifier can be estimated on the basis of the classifier test scores.

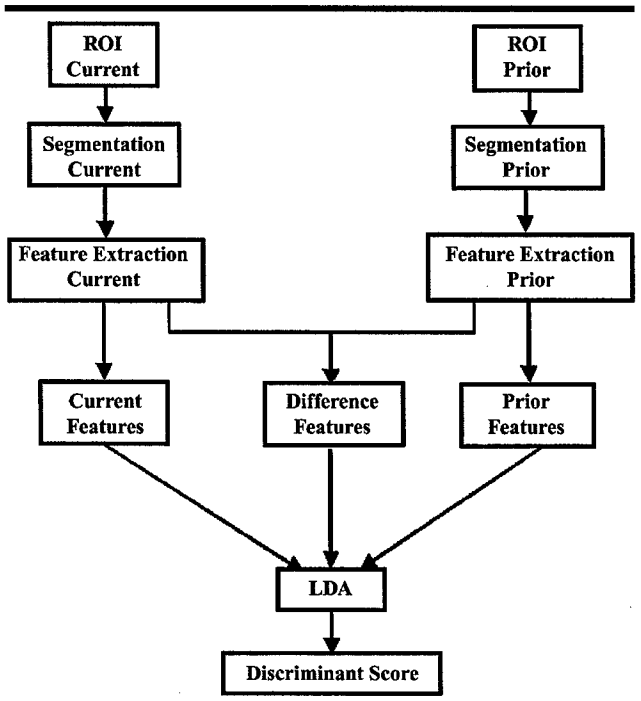


Figure 3. Flowchart of the classification method. *LDA* = linear discriminant classifier.

Relative Computer Malignancy Rating of the Masses

A relative computer malignancy rating ranging from 1 to 10 was provided to the radiologists for the reading with CAD. The relative malignancy rating was obtained by linearly scaling the classifier output within the interval between 1 and 10 and then rounding the result to the nearest integer. A rating of 1 corresponded to the highest score of the mass being benign, and a rating of 10 corresponded to the highest score of the mass being malignant. This transformation provided a more intuitive presentation of the scale to the observer than did the original classifier output. The linear transformation was not used to evaluate the classifier accuracy in terms of the class distributions or in terms of ROC analysis. Gaussian functions were fitted to the distributions of the malignant and benign samples to obtain a fitted binormal distribution with the classifier's malignancy ratings of 1–10 (Fig 4b). The fitted distribution was displayed on a graphical user interface as a reference when the radiologist evaluated the temporal pairs using CAD.

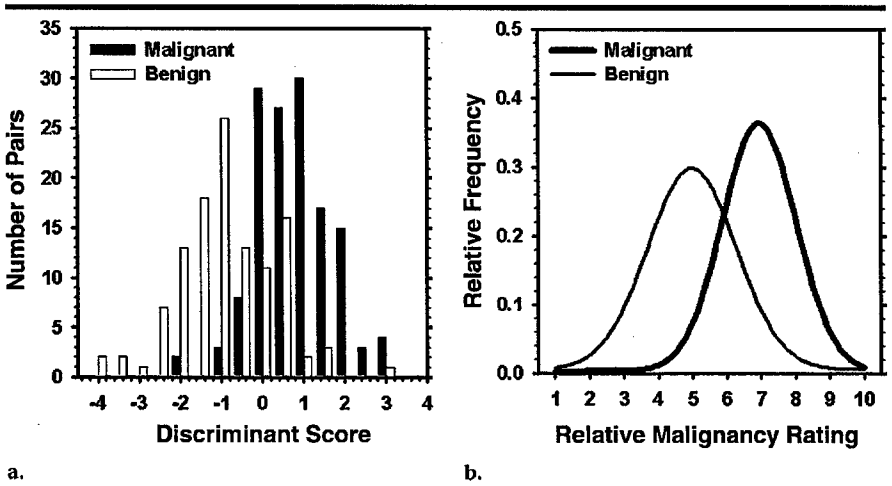


Figure 4. (a) Histogram of the classifier's test scores. (b) Binormal distribution fitted to the histogram of the classifier's test scores.

Observer Performance Study

The observer study evaluated the radiologist's performance in the classification of malignant and benign breast masses by interpreting a temporal pair of ROIs containing the mass on a display monitor. The radiologist was asked to provide an estimate of the likelihood of malignancy by using a 0%–100% scale and the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) assessment (21) of each mass. The study was designed in two reading conditions. The first reading condition is referred to as the independent mode, in which the radiologist read the masses without computer aid. The second reading condition is referred to as the sequential mode, in which the radiologist initially read a temporal pair without computer aid and then read the same pair with computer aid. First, the ratings without computer aid were recorded and then the computer rating of the mass was displayed on the monitor. The radiologist recorded the final rating after taking into consideration the computer rating. For simplicity of presentation, we will consider that there are a total of three modes from the aforementioned two readings—independent mode, sequential mode without CAD, and sequential mode with CAD. The sequential mode without CAD differs from the independent mode only in that the reader knew that the computer information would immediately follow. Eight radiologists (A.N., C.B., C.P., D.A., J.B., K.K., M.A.R., and S.P.) approved by the Mammogra-

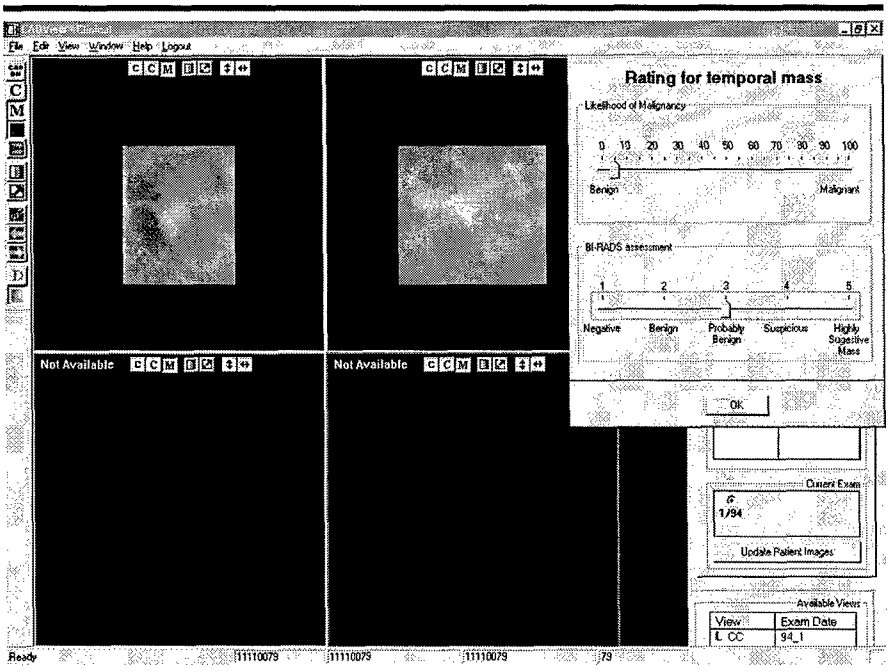


Figure 5. Example of graphical user interface shows reading in the independent mode or sequential mode without CAD. ROI on the left is prior and that on the right is current. The radiologist provided two ratings: an estimate of the likelihood of malignancy and the BI-RADS assessment, shown at the upper right area of the screen.

phy Quality Standards Act and two breast imaging fellows (M.F. and J.S.) participated as observers in this study. (There was no correspondence between order of the observers above and the observers' numeric order in the Results section and the Tables.) The eight radiologists had experience in mammography that ranged from 3 to 24 years. The breast imaging fellows were certified by the American Board of Radiology and had at least 3 months of experience in breast imaging.

For the observer experiments, the 253 pairs of images containing masses were divided into four non-overlapping groups, with approximately one-quarter of the pairs in each group. Each radiologist participated in four reading sessions. In each session, the observer read the pairs of images of one group in independent mode and those of another group in sequential mode so that no pairs of images would be read in both modes in a single session. The reading order of the temporal pairs of images within one group was randomized for each observer. Each observer would read in the independent mode first and then in the sequential mode in two of the sessions and vice versa in the other two sessions. We systematically arranged the reading order of the groups and the order of the modes to balance the frequency of both in the reading sessions. This counterbalanced design was intended to minimize the potential effects such as learning, fatigue, and memorization on the outcomes of the observer experiments. For each radiologist, there was at least a 1-month interval between reading pairs of images of the first two groups and those of the second two groups to avoid recall bias. All 10 observers read the temporal pairs independently.

Each observer underwent a training session in which the purpose of the study, the experimental procedure, the rating scales, the performance of the computer classifier, and the computer's rating scale were explained. The observer was also informed that the pairs of images included normal tissues in addition to malignant and benign masses. The prevalence of the malignant masses in the data set was not disclosed to the observer either in the training session or in the actual reading session. The observer then read 10 temporal pairs of images that were not used in the actual experiments to familiarize the observer with the reading processes and the user interface. The observer was informed of the true pathologic findings after rating each training case so that the findings could be compared to the observer's own ratings and the computer rating. However, in the actual experiment, no information regarding the true findings was provided after the readings.

A graphical user interface was developed to present the temporal pairs of images containing ROIs to the radiologists (Figs 5, 6). The observer assessed the two ROIs of a temporal pair that were displayed side-by-side on a display workstation. The observers provided estimates of the likelihood of malignancy by using a scale of 1%–100% and by choosing one of the five standard BI-RADS categories: negative, benign, probably benign, suspicious, and highly suggestive of malignancy. When the computer rating was displayed in the sequential mode with CAD, the fitted binomial distribution of the relative computer malignancy rating was presented to the radiologists (Fig 4b) as a reference. The radiologists were allowed unlimited time for the evaluation of the temporal pairs. For each radiologist, we recorded the time for the evaluation of the temporal pairs in both independent and sequential modes.

Statistical Analysis

The likelihood of malignancy ratings by the individual observers for the different reading conditions was analyzed by using ROC methods. The classification accuracy was quantified by using the total A_z, as well as the partial area index (22) calculated above a sensitivity threshold of 0.90 (hereafter, ${}_{0.90}A_z'$). The A_z was estimated by using the Dorfman-Berbaum-Metz method for analysis of multireader multicase data (23), in which the maximum likelihood estimation of the binormal distributions was fitted to the observer ratings, deriving the ROC curve. The statistical significance of the difference in A_z between the different reading conditions was also estimated by using the Dorfman-Berbaum-Metz method, the Student paired t test for analysis of observer-specific paired data, and the Obuchowski method (24). The Obuchowski method, which was also generalized by Lee and Rosner (25) for multiple readers multiple modalities studies, accounts for the possible correlations that exist among the temporal pairs of images, such as craniocaudal and mediolateral oblique pairs in the same patient or pairs obtained from multiple years in the same patient.

The radiologists' diagnostic decision based on the BI-RADS assessment was analyzed in this study by partitioning the BI-RADS categories into two groups. Group 1 consisted of BI-RADS categories 1 and 2, and group 2 consisted of BI-RADS categories 3, 4, and 5. BI-RADS category 0 was not allowed. This partition-

ing was associated with the estimation of callbacks, referred to as the callback grouping. If a mass was assigned to group 1, then it was assumed that no callback would be recommended. If a mass was assigned to group 2, then it was assumed that at least callback would be recommended. Each of the temporal pairs of images for an observer reading in a given mode was then classified to be a member of one of the two groups on the basis of the BI-RADS assessment. The changes in the group membership for the temporal pairs were then tallied for the different modes. A second partitioning was performed by combining BI-RADS categories 1, 2, and 3 into group 1 and BI-RADS categories 4 and 5 into group 2. This partitioning was associated with the estimation of biopsy recommendations, referred to as the biopsy recommendation grouping. If a mass was assigned to group 1 then it was estimated that no biopsy would be recommended. If a mass was assigned to group 2 then it was assumed that biopsy would be recommended.

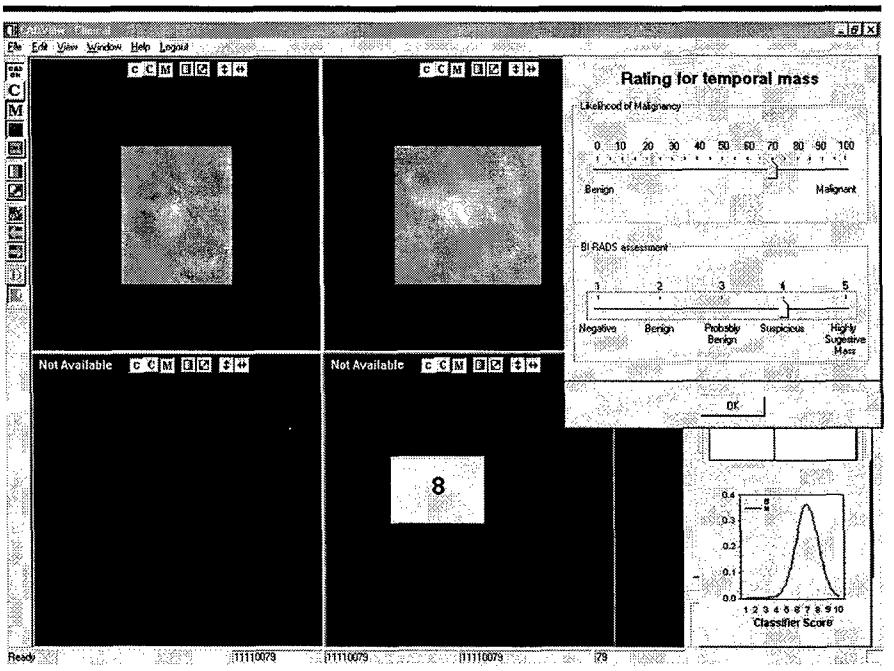


Figure 6. Example of a graphical user interface shows reading in the sequential mode with computer aid. ROI on the left is prior and that on the right is current. The computer rating (*8* in this example) was shown in the lower middle part of the screen. The performance of the computer classifier in terms of the distribution of the relative malignancy rating was shown in the lower right corner of the screen. In the sequential mode, the radiologist first evaluated the mass without CAD and then could change the likelihood of malignancy and/or BI-RADS assessment after taking into consideration the computer's rating.

RESULTS

The A_z values for the 10 radiologists participating in the study for the three reading modes are presented in Table 1. The average ROC curves for the three reading modes and the classifier are shown in Figure 7. The computer classifier's A_z value for the 253 temporal pairs was 0.87. The average ROC curves for the observers were obtained by averaging the fitted *a* and *b* parameters of the individual radiologist's ROC curve for each mode and then calculating the ROC curve. The A_z value was 0.79 for the independent mode, 0.81 for the sequential mode without CAD, and 0.84 for the sequential mode with CAD. The performance of the radiologist therefore improved, on average, when reading was made with computer aid. The improvement between the sequential mode with CAD and the independent mode was statistically significant (Table 2) ($P = .005$, Student paired *t* test; $P = .005$, Dorfman-Berbaum-Metz method; $P = .01$, Obuchowski method). In addition, the improvement in performance between the sequential mode with CAD and the sequential mode without CAD was also statistically significant ($P = .001$, Student paired *t* test; $P = .001$, Dorfman-Berbaum-Metz method; $P < .001$, Obuchowski method). An improvement was observed between the sequential mode without CAD and the independent mode, but it did not achieve statistical significance ($P = .137$, Student paired *t* test; $P = .139$, Dorfman-Berbaum-Metz method; $P = .073$, Obuchowski method).

TABLE 1
A_z Values for the Characterization of Masses in the Three Modes

Radiologist No.	Independent Mode A_z	Sequential Mode without CAD A_z	Sequential Mode with CAD A_z
1	0.84 ± 0.02	0.84 ± 0.02	0.89 ± 0.02
2	0.82 ± 0.03	0.84 ± 0.02	0.86 ± 0.02
3	0.75 ± 0.03	0.80 ± 0.03	0.81 ± 0.03
4	0.82 ± 0.03	0.88 ± 0.02	0.92 ± 0.02
5	0.74 ± 0.03	0.75 ± 0.03	0.80 ± 0.03
6	0.77 ± 0.03	0.85 ± 0.02	0.87 ± 0.02
7	0.76 ± 0.03	0.82 ± 0.03	0.83 ± 0.03
8	0.82 ± 0.03	0.76 ± 0.03	0.76 ± 0.03
9	0.75 ± 0.03	0.78 ± 0.03	0.83 ± 0.03
10	0.77 ± 0.03	0.74 ± 0.03	0.79 ± 0.03

Note.—Data are the mean ± standard deviation. The A_z value from the average *a* and *b* parameters was 0.79 for the independent mode, 0.81 for the sequential mode without CAD, and 0.84 for the sequential mode with CAD.

The computer classifier's A_z value was higher than the individual radiologists' A_z value obtained in the independent mode without CAD. In the sequential mode with CAD, radiologists 1 and 4 achieved higher A_z than did the computer classifier. Radiologist 6, who read in the sequential mode with CAD, obtained an A_z value of 0.87, which was the same as that of the computer classifier. The performance of radiologist 8 declined with the use of CAD; the A_z value was 0.82 for the independent mode and decreased to 0.76 for the sequential mode with CAD. However, when the sequential mode without CAD was compared with the sequential mode with CAD for this radiologist, there was no change ($A_z = 0.76$). For the rest of the radiologists, the improvement in A_z value ranged between 0.02 and 0.10.

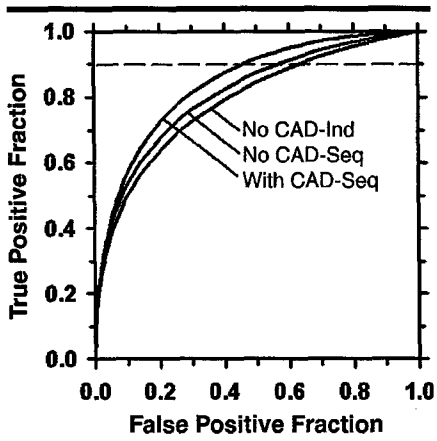


Figure 7. Average ROC curves for the three reading modes: independent (*No CAD-Ind*) (A_z = 0.79), sequential without CAD (*No CAD-Seq*) (A_z = 0.81), and sequential with CAD (*With CAD-Seq*) (A_z = 0.84).

Similar trends can be observed in the ${}_{0.90}A_z'$ values for the three reading modes (Table 3). The computer classifier's ${}_{0.90}A_z'$ value was 0.52. The statistical significance of the differences between every two of the three modes is presented in Table 4. The improvement in the radiologists' classification accuracy for the sequential mode with CAD (${}_{0.90}A_z'$ = 0.37) compared with that for the independent mode (${}_{0.90}A_z'$ = 0.21) was statistically significant (P = .005, Student paired t test). Similarly, the improvement for the sequential mode with CAD (${}_{0.90}A_z'$ = 0.37) compared with that for the sequential mode without CAD (${}_{0.90}A_z'$ = 0.26) was also statistically significant (P = .001, Student paired t test). Again, an improvement was observed between the sequential mode without CAD and the independent mode, but it did not achieve statistical significance (P = .180, Student paired t test). For radiologist 8, there was an improvement in the ${}_{0.90}A_z'$ value for the readings in the sequential mode without CAD (${}_{0.90}A_z'$ = 0.14) and then with use of CAD (${}_{0.90}A_z'$ = 0.17).

On the basis of the BI-RADS assessment, the computer classifier's influence on the radiologist's diagnostic decision was evaluated. Results of the callback grouping based on the BI-RADS assessment for the three modes are presented in Table 5. When the radiologists evaluated the temporal pairs in the sequential mode with CAD, an average (per radiologist) of 2.3% (3.2 of 138) of additional malignant masses were correctly recommended for callback and 0.6% (0.7 of 115) of additional benign masses were incorrectly recommended for callback compared with the evaluation in the independent mode. The reading in sequential mode with CAD compared with reading in sequential mode without CAD resulted in an average of 1.4% (1.9 of 138) of additional correct callbacks for malignant masses and 2.1% (2.4 of 115) of additional incorrect callbacks for benign masses. A comparison of the results obtained from readings in the independent mode and the sequential mode without CAD is shown also in Table 5. Although both readings were conducted without CAD, there was, on average, a correct reduction of 1.5% (1.7 of 115) of callbacks for benign masses and an increase of 0.9% (1.3 of 138) of callbacks for malignant masses when reading in the sequential mode without CAD.

Results of the biopsy recommendation grouping based on BI-RADS assessment for the three modes are presented in Table 6. When the radiologists evaluated the temporal pairs in the sequential mode with CAD compared with the independent mode, an improvement was obtained. There was an average reduction of 0.7% (0.8 of 115) of recommended biopsies for benign masses, and an additional 5.7% (7.8 of 138) of malignant masses were correctly recommended for biopsy. A comparison of the sequential mode with CAD with the sequential mode without CAD indicated that an average of 1.0% (1.1 of 115) of additional incorrect biopsy recommendations for benign masses were made, and an additional 4.0% (5.5 of 138) of correct biopsy recommendations for malignant masses were made. Table 6 also shows the results of comparison of reading in the independent mode with reading in the sequential mode without CAD. On average, a 1.7% (1.9 of 115) of correct reduction in biopsy recommendations were observed for benign masses, and an additional 1.7% (2.3 of 138) of correct biopsy recommendations for malignant masses were made when reading was performed in the sequential mode without CAD.

The reading time per temporal pair of mammograms for the 10 radiologists was 8.3–26.0 seconds (mean, 14.5 seconds) in the independent mode and 13.8–35.8 seconds (mean, 18.8 seconds) in the sequential mode. The increase in the reading time in the sequential mode (reading without CAD followed by reading with CAD) compared with reading in the independent mode (reading only without CAD) was statistically significant (P < .001).

DISCUSSION

In this ROC study, we observed an improvement in the radiologists' performance in the estimation of the likelihood of malignancy of masses seen on temporal pairs of mammograms when the radiologists read with computer aid. An improvement was also observed when the performance was evaluated in terms of BI-RADS assessment. To our knowledge, this was the first ROC study in which masses were evaluated by the radiologists on temporal pairs of mammograms and the computer classifier also used information regarding the temporal change in the classification of masses.

To study if the presence of a computer influences observer performance, we have used two reading modes without CAD: the independent mode and the sequential mode without CAD. We observed an interesting phenomenon: Seven of the 10 radiologists improved their A_z values when reading in the sequential mode without CAD. Although the improvement did not achieve statistical significance and intraobserver variability might have contributed to the differences, this appeared to be consistent with our observation in another clinical observer study (26) for breast cancer detection. In that study, the callback rate for the study group increased during the reading without CAD compared with that for the screening population not participating in the study, and the sensitivity of cancer detection was relatively high (91%) compared with the sensitivities reported in the literature. This reflects the possibility of a subtle change in the behavior when that behavior is being studied.

In two other observer studies, in which the effect of CAD on radiologists' performance in detection of lung nodules was evaluated (27,28), the independent and sequential modes were also compared. Kobayashi et al (27) found that 10 of 16 observers improved their A_z values when reading in the sequential mode without CAD compared with reading in independent mode. The average A_z value for the 16 observers was 0.894 for the independent mode and 0.906 for the sequential mode without CAD. In another study (28), the average A_z value for the independent mode was 0.829 and that for sequential mode without CAD was 0.835. Therefore, in both studies, the same trend was observed as in our studies, although the differences again did not achieve statistical significance.

Beiden et al (29) discussed the psychologic phenomenon of reader vigilance even though it did not show statistically significant change in the radiologists' performance in the aforementioned studies. Many radiologists may operate at a higher sensitivity level if they are aware that their performance is being evaluated. This awareness is accentuated when the computer's reading is displayed immediately after the radiologist's reading of each temporal pair of mammograms. There are exceptions. In our study, the performance of two of the radiologists (radiologist 8 and radiologist 10) decreased when the independent reading and the sequential reading without CAD were compared. However, if we compared the readings in the sequential mode without CAD and then with use of CAD, radiologist 8 showed an improvement in the $_{0.90}A'_z$ value. With CAD, radiologist 10 showed improved results, exceeding that of the reading in the independent mode.

The performance in terms of A_z and $_{0.90}A'_z$ values was better in sequential mode with CAD than in the other modes. The improvement between reading in the sequential mode with CAD (A_z = 0.84, $_{0.90}A'_z$ = 0.37) and the independent mode (A_z = 0.79, $_{0.90}A'_z$ = 0.21) was greater than the improvement between reading in the sequential mode with CAD (A_z = 0.84, $_{0.90}A'_z$ = 0.37) and the sequential mode without CAD (A_z = 0.81, $_{0.90}A'_z$ = 0.26). However, reading in the sequential mode with CAD versus the sequential mode without CAD had higher statistical significance (P = .001, Student paired t test; P = .001, Dorfman-Berbaum-Metz method; P < .001, Obuchowski method for A_z difference; P = .001, Student paired t test for $_{0.90}A'_z$ difference) than reading in the sequential mode with CAD versus the independent mode (P = .005, Student paired t test; P = .005, Dorfman-Berbaum-Metz method; P = .01, Obuchowski method for A_z difference; P = .005, Student paired t test for $_{0.90}A'_z$ difference). This finding may be attributed to the fact that the correlation between the scores in the sequential mode with and without CAD is higher than the correlation between the scores in the independent mode and the sequential mode with CAD. The higher correlation leads to a smaller variance for the difference between reading in the sequential mode with and without CAD and thus a higher statistical significance in their difference.

Beiden et al (29) analyzed the variance components of the ROC accuracy measures for comparing independent versus sequential reading and reached the conclusion that sequential reading is expected to achieve higher statistical significance. Our results appear to be consistent with this expectation. The estimation based on the Obuchowski analysis that accounted for the possible correlation among the pairs of images did not change the trend or statistical significance of the results in comparison with those obtained with 253 temporal pairs of images.

The BI-RADS assessments provided by the radiologists allowed an estimation of the specific action that the radiologists would take after evaluating the temporal pairs of images. Generally, when the radiologists used CAD, they correctly recommended additional callbacks for malignant masses but also increased the callbacks for benign masses. This indicates that the radiologists would increase their sensitivity but might also reduce their specificity when they used CAD as discussed earlier and by Helvie et al (26). However, when the independent mode is compared with the sequential mode without CAD in terms of callback, we again observe the phenomenon that the radiologists were influenced by the presence of the computer. In this case, the trend is different: On average, the radiologists had a slight decrease in callbacks for benign masses and a correct increase in callbacks for malignant masses when evaluating in the sequential mode without CAD.

TABLE 2
Statistical Significance of the Difference in A_z Values for the Three Modes

Comparison	Independent Mode A_z	Sequential Mode without CAD A_z	Sequential Mode with CAD A_z	Student Paired t Test	Dorfman-Berbaum-Metz Method
1	0.79	0.81	Not applicable	P = .137	P = .139
2	0.79	Not applicable	0.84	P = .005*	P = .005*
3	Not applicable	0.81	0.84	P = .001*	P = .001*

* Statistically significant difference at P < .05 level.

TABLE 3
$_{0.90}A'_z$ Values for the Characterization of Masses in the Three Modes

Radiologist No.	Independent Mode $_{0.90}A'_z$	Sequential Mode without CAD $_{0.90}A'_z$	Sequential Mode with CAD $_{0.90}A'_z$
1	0.36 ± 0.06	0.36 ± 0.06	0.55 ± 0.06
2	0.32 ± 0.07	0.33 ± 0.07	0.43 ± 0.07
3	0.13 ± 0.05	0.22 ± 0.07	0.27 ± 0.07
4	0.22 ± 0.06	0.41 ± 0.08	0.61 ± 0.06
5	0.15 ± 0.05	0.12 ± 0.04	0.25 ± 0.06
6	0.13 ± 0.05	0.41 ± 0.07	0.43 ± 0.07
7	0.17 ± 0.05	0.20 ± 0.06	0.22 ± 0.07
8	0.25 ± 0.06	0.14 ± 0.04	0.17 ± 0.05
9	0.12 ± 0.04	0.19 ± 0.06	0.30 ± 0.06
10	0.20 ± 0.05	0.20 ± 0.05	0.34 ± 0.06

Note.—Data are the mean ± standard deviation. The A_z value from the average a and b parameters was 0.21 for the independent mode, 0.26 for the sequential mode without CAD, and 0.37 for the sequential mode with CAD.

TABLE 4
Statistical Significance of the Difference in $_{0.90}A'_z$ Values for the Three Modes

Comparison	Independent Mode $_{0.90}A'_z$	Sequential Mode without CAD $_{0.90}A'_z$	Sequential Mode with CAD $_{0.90}A'_z$	Student Paired t Test
1	0.21	0.26	Not applicable	P = .180
2	0.21	Not applicable	0.37	P = .005*
3	Not applicable	0.26	0.37	P = .001*

* Statistically significant difference at P < .05 level.

TABLE 5
Results of the Callback Grouping Based on BI-RADS Assessment for the Three Modes

	Sequential Mode with CAD vs Independent Mode		Sequential Mode with CAD vs Sequential Mode without CAD		Sequential Mode without CAD vs Independent Mode	
Radiologist No.	Change in Callbacks for Benign Masses*	Change in Callbacks for Malignant Masses†	Change in Callbacks for Benign Masses*	Change in Callbacks for Malignant Masses†	Change in Callbacks for Benign Masses*	Change in Callbacks for Malignant Masses†
1	+4	+4	+1	+1	+3	+3
2	−6	+5	0	+2	−6	+3
3	−2	+10	0	+1	−2	+9
4	+10	+7	+11	+3	−1	+4
5	+8	+5	+6	+8	+2	−3
6	−10	+4	+1	0	−11	+4
7	−2	+1	+1	0	−3	+1
8	+3	−5	0	0	+3	−5
9	+4	−1	+1	0	+3	−1
10	−2	+2	+3	+4	−5	−2
Total	+7	+32	+24	+19	−17	13
Average ± SD per Radiologist	+0.7 ± 6.2	+3.2 ± 4.2	+2.4 ± 3.5	+1.9 ± 2.6	−1.7 ± 4.7	+1.3 ± 4.1
Average ± SD (%)‡	+0.6 ± 5.3	+2.3 ± 3.0	+2.1 ± 3.0	+1.4 ± 1.9	−1.5 ± 4.1	+0.9 ± 3.0

Note.—Unless otherwise specified, data are the number of callbacks. SD = standard deviation.
* Positive numbers indicate incorrect increase in callbacks. Negative numbers indicate correct decrease in callbacks.
† Positive numbers indicate correct increase in callbacks. Negative numbers indicate incorrect decrease in callbacks.
‡ Obtained by dividing the average per radiologist by 115 for benign masses and by 138 for malignant masses.

TABLE 6
Results of the Biopsy Recommendation Grouping Based on BI-RADS Assessment for the Three Modes

	Sequential Mode with CAD vs Independent Mode		Sequential Mode with CAD vs Sequential Mode without CAD		Sequential Mode without CAD vs Independent Mode	
Radiologist No.	Change in Biopsy Recommendation for Benign Masses*	Change in Biopsy Recommendation for Malignant Masses†	Change in Biopsy Recommendation for Benign Masses*	Change in Biopsy Recommendation for Malignant Masses†	Change in Biopsy Recommendation for Benign Masses*	Change in Biopsy Recommendation for Malignant Masses†
1	+2	+2	−3	−1	+5	+3
2	+2	+12	0	+8	+2	+4
3	+3	+6	−1	+2	+4	+4
4	+13	+10	+10	+3	+3	+7
5	−12	+4	−1	+10	−11	−6
6	−11	+11	−4	0	−7	+11
7	−5	+4	+1	0	−6	+4
8	+5	−1	0	+4	+5	−5
9	0	+13	+2	+8	−2	+5
10	−5	+17	+7	+21	−12	−4
Total	−8	+78	+11	+55	−19	+23
Average ± SD per Radiologist	−0.8 ± 7.6	+7.8 ± 5.7	+1.1 ± 4.3	+5.5 ± 6.6	−1.9 ± 6.6	+2.3 ± 5.5
Average ± SD (%)‡	−0.7 ± 6.6	+5.7 ± 4.1	+1.0 ± 3.7	+4.0 ± 4.8	−1.7 ± 5.7	+1.7 ± 4.0

Note.—Unless otherwise specified, data are the number of biopsy recommendations. SD = standard deviation.
* Positive numbers indicate incorrect increase in biopsy recommendation. Negative numbers indicate correct decrease in biopsy recommendation.
† Positive numbers indicate correct increase in biopsy recommendation. Negative numbers indicated incorrect decrease in biopsy recommendation.
‡ Obtained by dividing the average per radiologist by 115 for benign masses and by 138 for malignant masses.

Performance based on the estimation of biopsy recommendations was better for sequential mode with CAD than for the other two modes. We observed, on average, a correct decrease in biopsy recommendation for benign masses (0.7%, 0.8 of 115) and an increase in biopsy recommendation for malignant masses (5.7%, 7.8 of 138) in the sequential mode with CAD than in the independent mode. For sequential mode with CAD compared with sequential mode without CAD, the radiologists also achieved, on average, a correct increase in biopsy recommendation for malignant masses (4.0%, 5.5 of 138); however, they also incorrectly increased biopsy recommendation for benign masses (1.0%, 1.1 of 115). Again, it is possible to conclude that the radiologists operated in a higher sensitivity mode when they used CAD. In this case, they correctly increased, on average, the recommendation for biopsy of malignant masses and did not substantially increase the recommendation for biopsy of benign masses. Note that the ROC curve for the radiologists' reading with CAD is higher than the ROC curves for reading without CAD. The increase in sensitivity is therefore not a result of

changing the operating point along their ROC curve but an actual increase in their overall accuracy.

When we compare the independent mode with the sequential mode without CAD in terms of biopsy recommendation, the radiologists were influenced by the presence of the computer. On average, the radiologists correctly reduced biopsy recommendation for benign masses and increased biopsy recommendation for malignant masses. If the individual radiologist's decisions are reviewed, it can be seen that there are large variations regarding the effect of CAD. These variations may be caused by the differences in the radiologist's confidence levels in the CAD system. The positive effect may increase if the accuracy of the computer classifier is further improved or if the confidence of the radiologists increases after they accumulate more experiences in working with CAD.

The increase in the reading time for the sequential mode compared with that in the independent mode is owing to the fact that in the sequential mode two conditions were evaluated, reading without CAD followed by reading with CAD; whereas in the independent mode, only one condition was evaluated, reading without CAD. We did not observe correlation between the reading time and the observer performance results.

We did not observe a specific trend in the performance of the breast imaging fellows and the radiologists. This probably may be explained by the fact that we included only two imaging fellows, which was insufficient to show a trend.

There are some limitations of our study. Ideally, the classifier should be developed on the basis of an independent data set and then applied to the data set used to evaluate the radiologist performance. However, we were limited in the size of the data set with temporal pairs collected for this study. A split of the data set would reduce the statistical power of the study. We used a "leave one case out" resampling method to develop and test our classifier with the same data set as that used for the observer performance study. The method is well established in the pattern recognition literature as a statistically valid technique for estimation of the classifier performance in an unknown population. The test scores of the classifier were presented to the radiologists in the observer study. Furthermore, the purpose of this study was not to measure the absolute performance of the radiologists in comparison with that of the classifier. Rather, our goal was to demonstrate that there is a relative improvement in radiologists' performance when they use a computer classifier that has a reasonable performance as a second opinion. We believe that the use of a different data set will not change the conclusions as long as the computer classifier has a reasonable performance.

In conclusion, we have performed an observer study to evaluate the effects of CAD on radiologists' characterization of masses on serial mammograms. The radiologists have significantly ($P = .005$) improved their performance when reading with computer aid was compared with reading without computer aid. Additional biopsies were correctly recommended for the malignant masses when reading with computer aid, and some biopsies of benign masses were reduced. These results suggest that CAD may be helpful in improving the accuracy of biopsy recommendations. Further studies are needed to determine if these improvements can be realized in clinical settings, where the prevalence of malignancy is much lower than that in an observer study.

APPENDIX

Features related to texture, morphology, and spiculation were extracted from each mass. The texture features were based on run-length statistics matrices (30). The run-length statistics matrices were computed from the images obtained with the rubber-band-straightening transform (11). The rubber-band-straightening transform maps a band of pixels surrounding the mass onto the cartesian plane (a rectangular region). The texture features were extracted from the vertical and horizontal gradient-magnitude images (11). Five texture measures, namely, short-run emphasis, long-run emphasis, gray-level nonuniformity, run-length nonuniformity, and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, for each ROI, a total of 20 features were calculated. The definition of the feature measures based on run-length statistics matrices can be found in the literature (30).

Morphological features were extracted from the automatically segmented mass shape. Five of the morphological features were based on the normalized radial length, defined as the euclidean distance from the object's centroid to each of its edge pixels, that is, the radial length, and normalized relative to the maximum radial length for the object (15). The following five features of normalized radial length were extracted: mean, standard deviation, entropy, area ratio, zero crossing count. In addition, the perimeter, area, circularity, rectangularity, contrast, perimeter-to-area ratio, and Fourier descriptor were extracted. The definitions of the morphological features can be found in the literature (20,31). Three of the morphological features (perimeter, area, and perimeter-to-area ratio) are related to the mass size and thus are feature descriptors of the mass size.

A spiculation measure was defined for each pixel on the mass border by using the statistics based on the directions of image gradients of pixels outside the mass border, relative to the normal direction to the mass border. The statistics were determined in a 90° sector centered about the normal at the border pixel and outside of the mass border (19,20). The spiculation measure for each border pixel was normalized to be between 0 and $\pi/2$, with $\pi/4$ indicating a random orientation of image gradients and larger values indicating a higher likelihood of spiculation. Three features were extracted from the spiculation measure. The first feature was the average of the spiculation measure for all pixels on the mass boundary. The second feature was the percentage of border pixels with a spiculation measure larger than $\pi/4$. The third feature was the average of the spiculation measure for pixels with a spiculation measure larger than $\pi/4$.

Acknowledgment: The authors are grateful to Charles E. Metz, PhD, for the use of the LABMRMC program.

References

1. Greenlee RT, Hill-Harmon MB, Murray T, Thun M. Cancer statistics, 2001. CA Cancer J Clin 2001; 51:15–36.
2. Zuckerman HC. The role of mammography in the diagnosis of breast cancer. In: Ariel IM, Cleary JB, eds. Breast cancer, diagnosis and treatment. New York, NY: McGraw-Hill, 1987; 152–172.
3. Tabar L, Dean PB. The control of breast cancer through mammography screening: what is the evidence? Radiol Clin North Am 1987; 25:993–1005.
4. Sickles EA. Mammographic features of 300 consecutive nonpalpable breast cancers. AJR Am J Roentgenol 1986; 146:661–663.
5. Kopans DB. The positive predictive value of mammography. AJR Am J Roentgenol 1992; 158:521–526.
6. Adler DD, Helvie MA. Mammographic biopsy recommendations. Curr Opin Radiol 1992; 4:123–129.
7. Bassett L, Shayestehfar B, Hirbawi I. Obtaining previous mammograms for comparison: usefulness and costs. AJR Am J Roentgenol 1994; 163:1083–1086.
8. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3183 consecutive cases. Radiology 1991; 179:463–468.
9. Burnside E, Sickles E, Sohlich R, Dee K. Differential value of comparison with previous examinations in diagnostic ver-

sus screening mammography. AJR Am J Roentgenol 2002; 179:1173–1177.
10. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study. Radiology 1999; 212:817–827.
11. Sahiner B, Chan H, Petrick N, Helvie M, Goodsitt M. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Med Phys 1998; 25:516–526.
12. Huo Z, Giger M, Vyborny C, Metz C. Breast cancer: effectiveness of computer-aided diagnosis—observer study with independent database of mammograms. Radiology 2002; 224:560–568.
13. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
14. Chan H, Sahiner B, Petrick N, et al. Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. Phys Med Biol 1997; 42: 549–567.
15. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computer-aided image analysis. IEEE Trans Med Imaging 1993; 12:664–669.
16. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA. Classification of malignant and benign masses based on hybrid ART2LDA approach. IEEE Trans Med Imaging 1999; 18:1178–1187.
17. Tourassi G, Markey M, Lo J, Floyd C. A neural network approach to breast cancer diagnosis as a constraint satisfaction problem. Med Phys 2001; 28:804–811.
18. Hadjiiski L, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan M. Analysis of temporal change of mammographic features: computer-aided classification of malignant and benign breast masses. Med Phys 2001; 28:2309–2317.
19. Sahiner B, Chan HP, Petrick N, Hadjiiski LM, Helvie MA, Paquerault S. Active contour models for segmentation and characterization of mammographic masses. In: Proceeding of the 5th International Workshop on Digital Mammography. Toronto, Canada. Madison, Wis: Medical Physics, 2001; 357–362.
20. Sahiner B, Chan H, Petrick N, Helvie M, Hadjiiski L. Improvement of mammographic mass characterization using spiculation measures and morphological features. Med Phys 2001; 28:1455–1465.
21. American College of Radiology. Breast imaging and data system atlas (BI-RADS atlas). Reston Va: American College of Radiolgy, 2003.
22. Jiang Y, Metz C, Nishikawa R. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745–750.
23. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol 1992; 27: 723–731.
24. Obuchowski N. Nonparametric analysis of clustered ROC curve data. Biometrics 1997; 53:567–578.
25. Lee ML, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics. J R Stat Soc C Appl Stat 2001; 50:337–344.
26. Helvie MA, Hadjiiski LM, Makariou E, Chan HP, Petrick N, Lo SB. A non-commercial CAD system for breast cancer detection on screening mammograms achieves high sensitivity: a pilot clinical trial (abstr). Radiology 2002; 225(P):459.
27. Kobayashi T, Xu X, MacMahon H, Metz C, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 1996; 199:843–848.
28. U.S Food and Drug Administration, Center for Devices and Radiological Health. Radiological devices advisory panel meeting, March 5, 2001: review of Deus RapidScreen. Available at *www.fda.gov/search/databases.html*. Accessed June 7, 2002.
29. Beiden S, Wagner R, Doi K, et al. Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of component of variance. Acad Radiol 2002; 9:1036–1043.
30. Galloway MM. Texture classification using gray level run lengths. Comput Graph Image Proc 1975; 4:172–179.
31. Petrick N, Chan H, Sahiner B, Helvie M. Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms. Med Phys 1999; 26:1642–1654.

Computerized nipple identification for multiple image analysis in computer-aided diagnosis

Chuan Zhou,[a] Heang-Ping Chan, Chintana Paramagul, Marilyn A. Roubidoux, Berkman Sahiner, Labomir M. Hadjiiski, and Nicholas Petrick[b]
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 28 May 2004; revised 27 July 2004; accepted for publication 10 August 2004; published 30 September 2004)

Correlation of information from multiple-view mammograms (e.g., MLO and CC views, bilateral views, or current and prior mammograms) can improve the performance of breast cancer diagnosis by radiologists or by computer. The nipple is a reliable and stable landmark on mammograms for the registration of multiple mammograms. However, accurate identification of nipple location on mammograms is challenging because of the variations in image quality and in the nipple projections, resulting in some nipples being nearly invisible on the mammograms. In this study, we developed a computerized method to automatically identify the nipple location on digitized mammograms. First, the breast boundary was obtained using a gradient-based boundary tracking algorithm, and then the gray level profiles along the inside and outside of the boundary were identified. A geometric convergence analysis was used to limit the nipple search to a region of the breast boundary. A two-stage nipple detection method was developed to identify the nipple location using the gray level information around the nipple, the geometric characteristics of nipple shapes, and the texture features of glandular tissue or ducts which converge toward the nipple. At the first stage, a rule-based method was designed to identify the nipple location by detecting significant changes of intensity along the gray level profiles inside and outside the breast boundary and the changes in the boundary direction. At the second stage, a texture orientation-field analysis was developed to estimate the nipple location based on the convergence of the texture pattern of glandular tissue or ducts towards the nipple. The nipple location was finally determined from the detected nipple candidates by a rule-based confidence analysis. In this study, 377 and 367 randomly selected digitized mammograms were used for training and testing the nipple detection algorithm, respectively. Two experienced radiologists identified the nipple locations which were used as the gold standard. In the training data set, 301 nipples were positively identified and were referred to as visible nipples. Seventy six nipples could not be positively identified and were referred to as invisible nipples. The radiologists provided their estimation of the nipple locations in the latter group for comparison with the computer estimates. The computerized method could detect 89.37% (269/301) of the visible nipples and 69.74% (53/76) of the invisible nipples within 1 cm of the gold standard. In the test data set, 298 and 69 of the nipples were classified as visible and invisible, respectively. 92.28% (275/298) of the visible nipples and 53.62% (37/69) of the invisible nipples were identified within 1 cm of the gold standard. The results demonstrate that the nipple locations on digitized mammograms can be accurately detected if they are visible and can be reasonably estimated if they are invisible. Automated nipple detection will be an important step towards multiple image analysis for CAD.
[DOI: 10.1118/1.1800713]

Key words: computer-aided detection, mammography, nipple detection, texture orientation field analysis

I. INTRODUCTION

Breast cancer is one of the leading causes for cancer mortality among women.[1] The most successful method for the early detection of breast cancer is screening mammography.[2,3] It has been demonstrated that an effective computer-aided diagnosis (CAD) system can provide a second opinion to the radiologists and improve the accuracy of detection and characterization of mammographic abnormalities, which, in turn, may reduce unnecessary biopsies. In clinical practice, radiologists routinely use a cranio-caudal (CC) and a mediolateral oblique (MLO) view along with mammograms obtained in previous years, for detecting and interpreting breast lesions. The multiple views allow for imaging of most of the breast tissue and increase the chance of the breast lesion to be detected. Our previous studies have demonstrated that computerized multiple view analysis could not only improve breast lesion detection with two-view information fusion,[4,5] but also improve malignant and benign lesion characterization by interval change analysis.[6] Our techniques used the nipple location, the only reliable landmark on the mammogram, as the reference point for two-view (CC and MLO views) information fusion and regional registration of tem-

poral pairs of mammograms of the same view. However, the nipple location was manually identified on the mammograms in these studies.

Automated methods for detection of the nipple location have been reported by Chandrasekhar,[7] Mendez,[8] and Yin.[9] In their methods, the breast boundary was extracted and then the nipple location was identified by searching for the maximum and minimum of the gradient changes or average intensity in a small region along the breast boundary. However, without mentioning the use of a training data set or how to train the detection program, Chandrasekhar *et al.* reported the performance of their method using a very limited data set of 24 images with 8 CC views and 16 oblique views. For 23 of the images (96%), the root-mean-square error of their detection method was reported to be less than 1 cm at an image resolution of 400 μm $\times$ 400 μm per pixel. Mendez *et al.* tested 156 mammograms that included lateral oblique and CC views. They reported that the average distance between the detected nipple location and the true position identified by two radiologists was 13.5 mm. Mendez *et al.* also tested Yin's method using the same 156 mammograms and obtained an average distance of 16.5 mm, while Yin *et al.* reported an average distance of 10 mm when tested on 80 mammograms. Neither Mendez *et al.* nor Yin *et al.* reported whether the nipple was in profile on the images, nor reported results for both training and test sets.

In a random sample of mammograms, many nipples cannot be positively identified, even by experienced mammography radiologists. Breast boundary-based methods therefore cannot accurately locate these nipples. For the cases that the nipple is not readily visible, a radiologist may examine the patterns of glandular tissue and ducts to find where they converge, and then estimate the nipple location in the convergent area. However, to our knowledge, no study has been reported to use texture convergence information for computerized nipple detection.

Computerized identification of nipple location on digitized mammograms is challenging because of the variations in image quality and in the nipple projections, especially for the nipples that are very flat and nearly invisible on the mammograms. In this study, we developed an automated technique for nipple identification on digitized mammograms with the information of nipple intensity changes, nipple geometric characteristics, and texture convergence toward nipple. Automated nipple detection will be the fundamental step towards the development of a multiple-image CAD system using our image registration techniques.

II. MATERIALS AND METHODS

A. Database

A total of 744 mammograms of 182 patients was used in our study. A data set consisting of 377 mammograms of 77 patients was used as training data set for development of the algorithms and 367 mammograms of 105 patients were used as the test data set. The mammograms were randomly selected from the patient files in the Department of Radiology at the University of Michigan with approval of the Institutional Review Board (IRB). The mammograms were acquired with GE mammography systems and were digitized with a LUMISYS 85 laser film scanner with a pixel size of 50 μm $\times$ 50 μm and 4096 gray levels. The gray levels are linearly proportional to optical densities (O.D.) from 0.1 to greater than 3 O.D. units. The nominal O.D. range of the scanner is 0–4. The full resolution mammograms were first smoothed with a 16 $\times$ 16 box filter and subsampled by a factor of 16, resulting in 800 μm $\times$ 800 μm images of approximately 225 $\times$ 300 pixels in size.

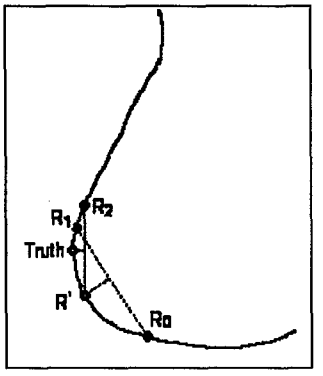


FIG. 1. Estimation of "gold standard" for invisible nipple images.

The 744 mammograms were randomly divided into a training and a test data set of 377 and 367 mammograms, respectively. For each mammogram, the image was first displayed on a monitor and visually inspected using windowing functions. According to the appearance of the nipple profile projection on the mammograms, the mammograms were classified into one of two classes: visible nipple class in which the nipple profiles were clearly projected on the mammogram and positively identifiable, and the invisible nipple class in which the nipple locations could not be positively identified. 301 of the 377 training images and 298 of the 367 test images were classified into the visible nipple class, while the remaining 76 and 69 images in the training and test data sets, respectively, were classified into the invisible nipple class.

In each mammogram, the nipple location was identified by experienced Mammography Quality Standards Act (MQSA) radiologists. This location was used as the "gold standard" for training the algorithms and evaluating of the computer performance. The radiologist visually inspected the image displayed on a monitor with a graphical user interface and used the windowing function to enhance the breast boundary. The radiologist marked the nipple location by using the cursor. One radiologist estimated the nipple location for all of the images in the visible nipple class. For the invisible nipple class, one radiologist estimated the nipple locations twice, another radiologist estimated the nipple location only once. The "gold standard" was estimated by averaging the radiologists' readings. Since the breast boundary is not a straight line, the averages of the x and y coordinates of two points along the breast boundary generally do not fall on the boundary. An average between two readings was thus estimated as the intersection between the breast

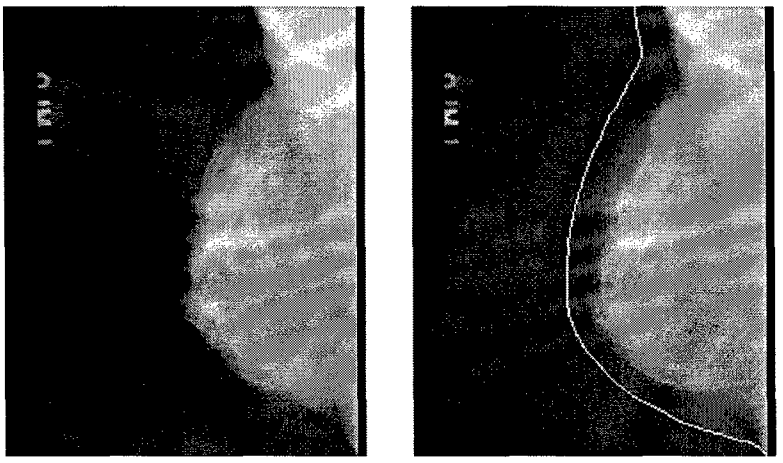


FIG. 2. (a) A mammogram from our image database; (b) the image superimposed with the detected breast boundary.

boundary and the normal to the midpoint of the line connecting the two readings, as shown in Fig. 1. When the two readings are not too far apart, this method is very close to that obtained by finding their midpoint along the breast boundary. However, this method is less prone to error if the breast boundary points are noisy. Using this averaging method, the average point R' was first estimated from Radiologist 1's two readings R_0 and R_1, then the "gold standard" was found as the average of point R' with Radiologist 2's reading R_2.

B. Breast boundary detection

The detection of breast boundary was the first step in our computerized nipple detection algorithm. The breast boundary separated the breast from the surrounding background which included the directly exposed area, the patient identification information, and lead markers. Computerized analysis was then performed only around the breast region after boundary detection. The breast boundary was first identified by a boundary tracking technique. The automated boundary tracking technique previously developed[10,11] was modified to improve its performance. The breast boundary was identified by a gradient-based method as follows. The background of the image was estimated initially by searching for the largest background peak from the gray level histogram of the image. A preliminary edge was found by a horizontal line-by-line gradient analysis starting from the top to the bottom of the image. The criterion used in detecting the edge points was the steepness of the gradient along the horizontal direction. The steeper the gradient, the greater the likelihood that an edge existed at that corresponding location. The preliminary edge served as a guide for a more accurate tracking algorithm that was subsequently applied. The tracking of the breast boundary started from approximately the middle of the breast image and moved upward and downward along the boundary. The direction to search for a new edge point was guided by the previously tracked edge points. The edge location was determined by searching for the maximum gradient along the gray level profile normal to the tracking direction. Because the boundary tracking was guided by the preliminary edge and the previously detected edge points, it could steer around the breast boundary and was less prone to diversion by noise and artifacts. After upward and downward tracking was finished, the tracked edges were smoothed to remove noisy fluctuations. A simple linear interpolation was used to connect the edge points so that a continuous breast boundary was found. An example of the tracked breast boundary is shown in Figs. 2(a) and 2(b).

C. Limiting the nipple search region

If the breast is properly positioned for imaging, almost all the nipples will be located along or close to the breast boundary. Our nipple search was performed within a small window of 9×9 pixels along the breast boundary, with the center of the search window located at the boundary point.

Defining a small search region along the breast boundary would reduce the chance that jagged breast borders from noise and artifacts would result in false positive nipple identification. We designed a geometric convergence analysis to estimate a nipple search region where the nipple would most likely be located. In an ideal situation, the nipple was located close to the boundary, approximately in the middle region of

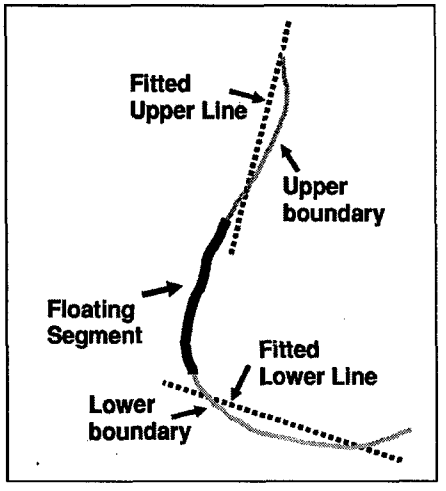


FIG. 3. Defining a limited nipple search region by geometric convergence analysis.

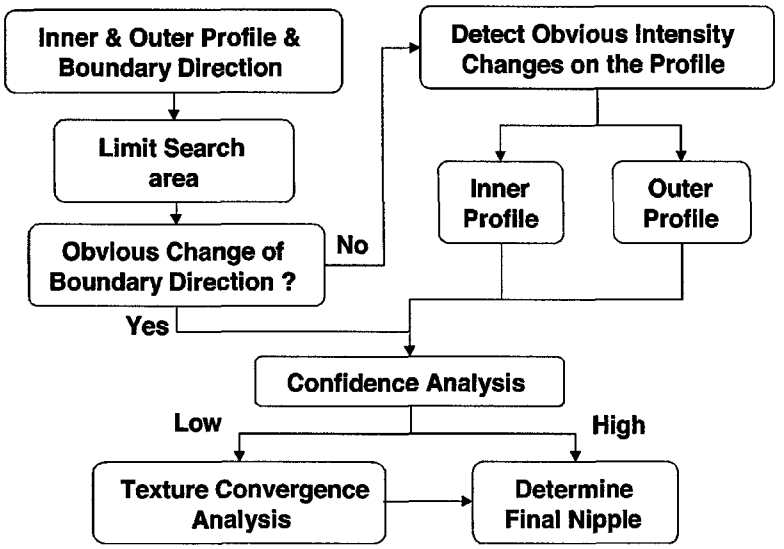


FIG. 4. Schematic of the automated nipple search method.

the breast for CC view and in the lower region for MLO view. As shown in Fig. 3, in the geometric convergence analysis, a floating segment containing 20% of boundary points was first placed at the middle of the breast boundary. The floating segment separated the boundary into an upper and a lower boundary segment. Two lines were then fitted to the boundary points in the upper and lower segments and the goodness-of-fit of the two lines was estimated by the sum of squares of the deviations between the fitted line and the boundary points. The convergence region was finally determined by moving the floating segment along the boundary until the deviation of the fitted lines from the breast boundary was minimized. The two fitted lines intersected the anterior region of the breast boundary at two points. The boundary region between these two points was defined as the nipple search region.

D. Nipple detection

1. Nipple search along breast boundary

After automated breast boundary detection, the breast boundary was smoothed to reduce small jagged fluctuations. From our analysis, we observed that there were sudden and distinct gray level changes in pixels close to the nipple for most of the mammograms with visible nipples. The direction of the breast boundary also had a sudden and distinct change when a convex nipple shape occurred along the breast boundary. In order to identify the location where these changes occurred, we constructed two smoothed intensity curves corresponding to the inner and outer intensity profiles and the curvature curve along the boundary, as defined in Eqs. (1)–(3). The curves were plotted against boundary point B_x, where $x=1,\ldots,n_B,n_B$ represented the total number of boundary points:

Inner intensity curve:

$$C_I(B_x)=\frac{1}{n_I}\sum_{k=1}^{n_I} f(k),\ k\in R_I,B_x\in \text{Breast Boundary}, \quad (1)$$

Outer intensity curve:

$$C_O(B_x)=\frac{1}{n_O}\sum_{k=1}^{n_O} f(k),\ k\in R_O,B_x\in \text{Breast Boundary}, \quad (2)$$

Curvature curve:

$$D(B_x)=\frac{1}{n_D}\sum_{k=1}^{n_D} d(k),\ k\in R_D,B_x\in \text{Breast Boundary}, \quad (3)$$

where R_I, R_O, and R_D were pixels within a 5×5 window of the inner profile, the outer profile, and 5 neighborhood boundary points, respectively. Each window was centered laterally at the current boundary point B_x. n_I, n_O, and n_D represented the number of pixels within each window. $f(k)$ and $d(k)$ were the gray level of the kth pixel within the window and the curvature at the kth boundary point, respectively. On the boundary point B_x, the first derivative, or the gradient, was estimated as the tangent T_x to the breast boundary at B_x. The curvature at B_x was the derivative of the gradient curve,[12] which represented the direction change of the tangent at boundary point B_x.

Figure 4 shows the nipple search scheme based on the boundary features. Nipple search was performed taking into account three situations in which the nipple exhibited different characteristics. First, a nipple shape was projected along the breast boundary. In the second and third situations, a nipple intensity profile could be identified inside or outside of the breast boundary. The details are described below.

Within the limited nipple search region, the first step was to detect if there was a sudden and distinct change in the boundary direction, which indicated a convex nipple shape outside of the boundary. The convex nipple could be detected by searching for the sharpest peak on the curvature curve. The peak feature p_R of every peak along the curve was calculated as the ratio of the peak height to the peak width. The sharpest peak was identified as the maximum of the peak features p_R. If the maximum peak feature p_R was larger than

a predefined threshold, then there was a convex nipple shape depicted on the boundary. The nipple location was identified as the peak point, N_{convex}, of the sharpest peak on the curvature curve. The threshold was determined using the training data set.

If no convex nipple could be found (i.e., no peak feature larger than the threshold), then the nipple search was performed by searching for obvious intensity changes along the inner and outer intensity profiles separately. Two peak features of the intensity curve were used to detect obvious intensity changes. The first peak feature p_R was estimated as the ratio of the peak height to the peak width. The second peak feature p_H was the peak height normalized to the sum of all the curve heights. If both p_R and p_H for a given peak were larger than the predefined thresholds, then it was an obvious intensity peak. The thresholds were again determined using the training data set. The most obvious intensity change was identified by the maximum p_H if more than one obvious intensity peaks were found along the intensity curve. If obvious intensity changes were found along both the inner and outer intensity curves, the potential nipple location $N_{intensity}$ was identified at the peak point of the intensity peak with maximum p_H on each curve. If the two maximum intensity peaks located on the inner and outer intensity curves were very close (defined as within 1 cm in our study), then the nipple location $N_{intensity}$ was identified to be the average of these two peak points. The average location was taken as the intercept of the breast boundary with the normal to the midpoint of the line connecting the two peak points. If these two peak points were not close, then the nipple location $N_{intensity}$ was determined as the maximum peak on the outer intensity profile because the outer intensity profile generally was less affected by structural noises. The nipple candidate N_{convex} or $N_{intensity}$ identified by this rule-based method is referred to as Nipple 1.

Due to the image quality, artifacts, or dense area near the breast border, the computer may identify a jagged breast boundary, which would lead to false detection of the nipple. To reduce the false detections, the identified candidate nipple location N_{convex} or $N_{intensity}$ was subjected to a confidence analysis. If there were several cosinelike peaks of similar size in the curvature curve or the inner intensity curve, it indicated that the breast boundary was jagged or there were dense tissues near the breast boundary, respectively. The confidence of the identified nipple was therefore set to low. The confidence was also set to low if the candidate nipple location N_{convex} or $N_{intensity}$ was null because the peak features were less than the predefined thresholds. In this situation, the nipple could not be found by the breast-boundary-based method described above and texture convergence analysis would be used as described next.

2. Nipple identification using texture convergence analysis

If the confidence of the rule-based nipple detection was set to low, a flow field based convergence analysis was initiated to estimate the nipple location based on the convergence of texture pattern of glandular tissues or ducts towards the nipple. The fibroglandular tissues or ducts appeared as oriented and flowlike textural pattern on the mammograms. With the assumption that there exists a dominant orientation at each pixel within a texture pattern, an "orientation image" can be computed from the gray level mammogram using least mean squares estimation based on Rao's optimal solution.[13] Let $g_x(u,v)$ and $g_y(u,v)$ represent the gradients at pixel (u,v) in the image. The gradient magnitude is computed as $G_{u,v}=\sqrt{g_x^2(u,v)+g_y^2(u,v)}$, and the gradient orientation is computed as $\theta_{u,v}=\arctan(g_y(u,v)/g_x(u,v))$. Assuming that the dominant orientation in a $N\times N$ local neighborhood centered at pixel (i,j) is $\phi(i,j)$, the sum-of-squares S can be computed as

$$S=\sum_{u=1}^{N}\sum_{v=1}^{N} G_{u,v}^2\cos^2(\theta_{u,v}-\phi(i,j)), \tag{4}$$

where S is the sum of the squared gradient magnitudes projected along a direction $\phi(i,j)$ in this neighborhood. $\phi(i,j)$ is the dominant orientation if S is the maximum. The maximum of S with respect to $\phi(i,j)$ can be found by solving the equation $(dS/d\phi(i,j))=0$,

$$\frac{dS}{d\phi(i,j)}=2\sum_{u=1}^{N}\sum_{v=1}^{N} G_{u,v}^2\cos(\theta_{u,v}-\phi(i,j))\sin(\theta_{u,v}-\phi(i,j)). \tag{5}$$

Thus, the dominant orientation $\phi(i,j)$ can be estimated as

$$\phi(i,j)=\frac{1}{2}\tan^{-1}\left(\frac{\sum_{u=1}^{N}\sum_{v=1}^{N} G_{u,v}^2\sin 2\theta_{u,v}}{\sum_{u=1}^{N}\sum_{v=1}^{N} G_{u,v}^2\cos 2\theta_{u,v}}\right)$$
$$=\frac{1}{2}\tan^{-1}\left(\frac{\sum_{u=1}^{N}\sum_{v=1}^{N} 2g_x(u,v)g_y(u,v)}{\sum_{u=1}^{N}\sum_{v=1}^{N}(g_x^2(u,v)-g_y^2(u,v))}\right). \tag{6}$$

Dense breasts generally exhibit more textural structures than fatty breasts on the mammograms. However, due to the presence of noise, the estimated local texture orientation may not always be correct. A low-pass filter can be used to find the local orientation that varies slowly in the local neighborhood. Before performing low-pass filtering, the orientation image was converted into a continuous vector field[13] defined as follows:

$$\Theta_x(i,j)=\cos(2\phi(i,j)) \tag{7}$$

and

$$\Theta_y(i,j)=\sin(2\phi(i,j)). \tag{8}$$

The low-pass filtering was performed by averaging of $\Theta_x(i,j)$ and $\Theta_y(i,j)$ in a local window with a size of 5×5 pixels, yielding $\Theta_x'(i,j)$ and $\Theta_y'(i,j)$, respectively, as the smoothed continuous vector field. The smoothed local orientation at (i,j) can then be computed as

$$O(i,j)=\frac{1}{2}\tan^{-1}\left(\frac{\Theta_y'(i,j)}{\Theta_x'(i,j)}\right). \tag{9}$$

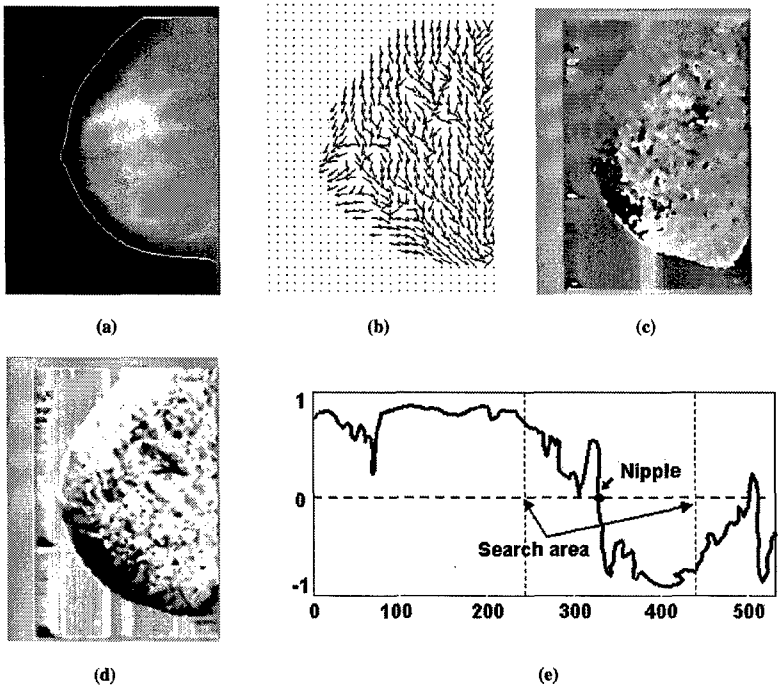


FIG. 5. An example of texture orientation field convergence analysis. (a) Original image superimposed with the detected breast boundary, (b) texture orientation field, (c) continuous orientation field O, (d) cosine component of continuous field O_x, (e) profile of O_x identified along the breast boundary.

Figure 5 shows an example of a computed orientation field superimposed on the original mammogram. The nipple location was indicated by the convergence of the estimated texture orientation. The following steps were used for the detection of the convergence of the texture orientation:

(1) Convert the smoothed orientation image into a continuous vector field:[14]

$$O_x(i,j) = \cos(O(i,j)). \tag{10}$$

(2) Identify the points of O_x in the inner profile region and then average to a 1D profile:

$$C_{Ox}(B_x) = \frac{1}{n_I}\sum_{k=1}^{n_I} O_x(k),\ k \in R_I, B_x \in \text{Breast Boundary}, \tag{11}$$

where n_I is the number of points within the local window represented by R_I. In our study, the size of R_I is 5 $\times$ 5. For simplicity, the index k is used to identify a point in R_I, replacing the indices (i,j).

(3) Detect the transition point of C_{Ox} by searching for the maximum gradient of C_{Ox} as shown in Fig. 5(e). A large gradient of the C_{Ox} indicated the convergence of the texture orientation which led to the location of nipple. A candidate nipple location O_x (Nipple2) was found if the maximum gradient was larger than a predefined threshold T_I. The threshold T_I was determined using the training data set.

In addition to the maximum gradient location, another indication of a nipple candidate is an approximately circular cluster of pixels with high orientation field strength. Because some of the nipples exhibited a convex shape, there would be a bright dot occupying several pixels on the image of orientation field O as shown in Fig. 5(c). Such bright dot indicated a candidate nipple, which we will refer to as Nipple3 in the following discussions. Note that, although the rule-based method could detect convex nipple location by searching for the maximum curvature of the breast boundary as described in Sec. II D 1, the confidence of the identified nipple might be set to low because of jagged breast boundary. In such a case, alternative nipple locations would need to be considered.

3. Determination of the final nipple location

After rule-based nipple detection along the boundary profile, and the convergence analysis using texture orientation field, three candidate nipple locations were obtained, as described above. Nipple1 was found by the rule-based method, Nipple2 was found by the change in the orientation projection O_x, and Nipple3 was found by the orientation field O. If the confidence of Nipple1 was set to high, the final nipple location was determined by Nipple1. Otherwise, the following rules were used to determine the final nipple location:

Situation 1: Both Nipple2 and Nipple3 could be detected by texture convergence analysis:

(1) If the distances between the three candidate nipples were all smaller than 0.5 cm (6 pixels), then the final nipple location was determined by Nipple1.
(2) If the distance between Nipple1 and Nipple2 was larger than 0.5 cm and the distance between Nipple1 and Nipple3 was smaller than 0.5 cm, then the final nipple location was determined by Nipple1.
(3) If the distance between Nipple1 and Nipple3 was larger than 0.5 cm and the distance between Nipple1 and Nipple2 was smaller than 0.5 cm, then the final nipple location was determined by Nipple1.

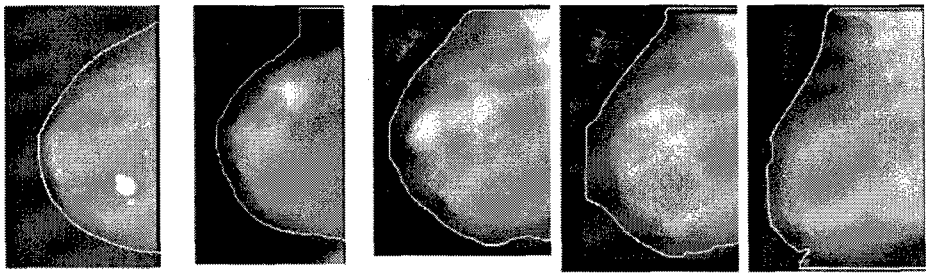


FIG. 6. Typical examples of tracked breast boundary rating: (a) rating as 0; (b) rating as 1+; (c) rating as 1−; (d) rating as 2−; (e) rating as 2+.

(4) If the distance between Nipple2 and Nipple3 was less than 0.5 cm but the distances from Nipple1 to both Nipple2 and Nipple3 were larger than 0.5 cm, then the final nipple location was determined by Nipple3.
(5) If the distances between every two of the three candidate nipples were larger than 0.5 cm, it indicated that the confidence of nipple detection using texture convergence analysis was low, then the final nipple location was determined by Nipple1. However, if Nipple3 was less than 0.5 cm from the breast boundary, it indicated that Nipple3 had higher confidence because nipple projection had a good convex shape, then Nipple3 was determined as the final nipple location.

Situation 2: only Nipple2 could be detected by texture convergence analysis:

(1) If the distance between Nipple2 and Nipple1 was smaller than 0.5 cm, then the final nipple location was determined by Nipple1.
(2) If the distance between Nipple2 and Nipple1 was larger than 0.5 cm, then the final nipple location was determined by Nipple2 if the maximum gradient of the 1D inner profile C_{Ox} of the smoothed continuous orientation field O was larger than another predefined threshold T_2 ($T_2 > T_1$); otherwise the final nipple location was determined by Nipple1. The threshold T_2 was determined using the training data set.

Situation 3: only Nipple3 could be detected by texture convergence analysis:

(1) Similar to Situation 2, if the distance between Nipple3 and Nipple1 was smaller than 0.5 cm, then the final nipple location was determined by Nipple1.
(2) If the distance between Nipple3 and Nipple1 was larger than 0.5 cm then the final nipple location was determined by Nipple3 if Nipple3 was less than 0.5 cm from the breast boundary; otherwise the final nipple was determined by Nipple1.

III. RESULTS

A. Breast boundary tracking

Our breast boundary tracking method was evaluated quantitatively in a previous study.[15] In this study, we applied the program to 744 mammograms. A qualitative performance evaluation of the tracked boundary was performed. Each of the computer tracked breast boundary was rated in three major categories and the "true" boundary was judged visually by an experienced medical physicist. If the boundary was very close to the true boundary it was rated as 0. Borders with a large section of local deviations were rated as 1− and 1+, where + and − indicated if the tracked boundary was outside or inside of the true boundary, respectively. Very poorly tracked borders or total failures were rated as 2− or 2+. Figure 6 shows typical examples of tracked breast boundary rating. The boundary shown in Fig. 6(a) is very closed to the true boundary and rated as 0. The upper section of the boundary was tracked outside the true boundary as shown in Fig. 6(b), which was rated as 1+. The lower section of the boundary was tracked into the breast region as shown in Fig. 6(c), which was rated as 1−. Figure 6(d) showed a very poorly tracked boundary that was rated as 2−. Figure 6(e) showed an example of failure in the lower part of the boundary tracking that tracked along the edge of the x-ray field and was rated as 2+. Of the 744 mammograms, 89.78% (668/744) of the tracked breast boundaries were rated as 0, 9.81% (73/744) were rated as 1+ or 1−, and 0.67% (5/744) were rated as 2− or 2+. The results showed that the boundaries in most of the mammograms in the data set were tracked very well. Although the boundaries which were rated as 1− and 1+ had local deviations, they were reasonably good to be used for nipple identification as discussed in Sec. IV.

TABLE I. Performance of the automated nipple detection program. The nipple detection accuracy is quantified as the percentage of images in which the detected nipple location is within 1 cm to the gold standard.

		Number of images	Rule-based method	Rule-based method with texture analysis
Training set	Visible	301	82.39% (248/301)	89.37% (269/301)
	Invisible	76	65.79% (50/76)	69.74% (53/76)
	All	377	79.05% (298/377)	85.41% (322/377)
Test set	Visible	298	89.93% (268/298)	92.28% (275/298)
	Invisible	69	47.83% (33/69)	53.62% (37/69)
	All	367	82.02% (301/367)	85.01% (312/367)

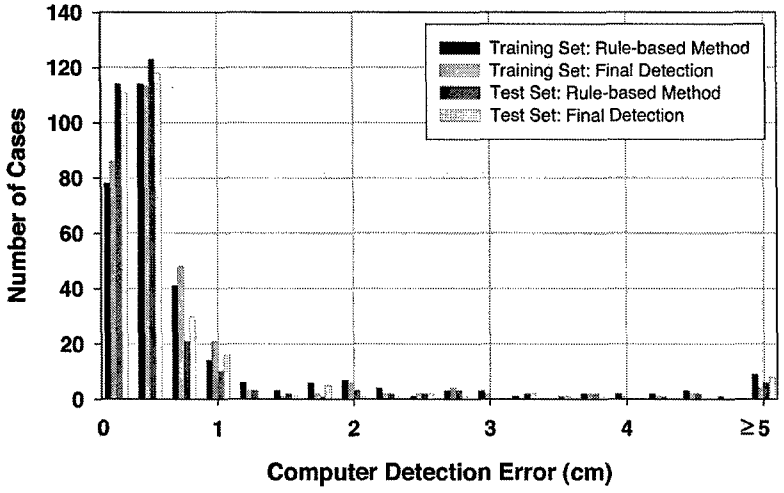


FIG. 7. Histogram of computer detection error (Euclidean distance from the detected nipple location to the "gold standard") for the mammograms in the visible nipple class.

B. Nipple identification

Because the diameters of nipples are larger than 1 cm for most patients, we chose the criterion of correct detection to be a distance of within 1 cm from the computer detected nipple location to the gold standard for evaluating the performance of the computerized nipple identification method. Table I shows the results for computer detected nipple location with an error within 1 cm of the gold standard. For the visible nipple images, the computer identified 89.37% (269/301) of the nipple location within 1 cm (mean =0.34 cm) of the gold standard for the training set, and 92.28% (275/298, mean=0.30 cm) of the nipple location within 1 cm of the gold standard for the test data set. For the invisible nipple images, the computer detected 69.74% (53/76, mean=0.24 cm) of the nipple location within 1 cm of the gold standard for the training data set, and 53.62% (37/69, mean=0.21 cm) of the nipple locations within 1 cm of the gold standard for the test set. The overall performance achieved by the computer in nipple detection including all images with visible or invisible nipple was 85.41% (322/377) and 85.01% (312/367) for the training and test data set, respectively.

To investigate the usefulness of the texture convergence analysis for nipple identification, we computed the nipple detection results without convergence analysis, in other words, by relying only on Nipple1 location identified by the rule-based method. This results in a simpler detection system, because none of the conditions in Sec. II D 3 are applied. In this situation, 82.39% (248/301) of the visible nipples and 65.79% (50/76) of the invisible nipples in the training data set, and 89.93% (268/298) of the visible nipples and 47.83% (33/69) of the invisible nipples in the test data set could be identified within 1 cm of the gold standard by using the rule-based nipple identification method. The mean errors under these conditions were 0.30 cm, 0.23 cm, 0.28 cm, and 0.18 cm, respectively. For all of the images including visible and invisible nipples, 79.05% (298/377) and 82.02% (301/367) of the nipple locations were identified within 1 cm of the gold standard. The images with errors larger than 1 cm were mainly caused by noise or artifacts along the breast boundary. Figures 7 and 8 show the histograms of the errors for our computerized nipple detection program for visible and invisible nipples, respectively. Figures 9 and 10 show the cumulative percentage of images

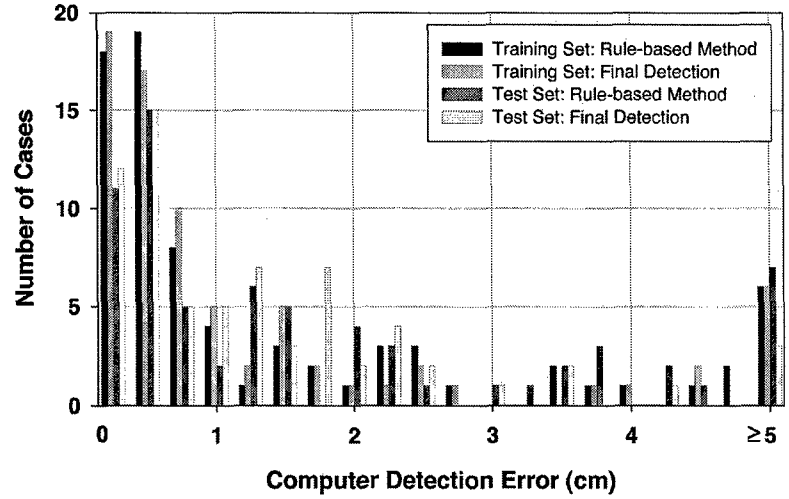


FIG. 8. Histogram of computer detection error (Euclidean distance from the detected nipple location to the "gold standard") for the mammograms in the invisible nipple class.

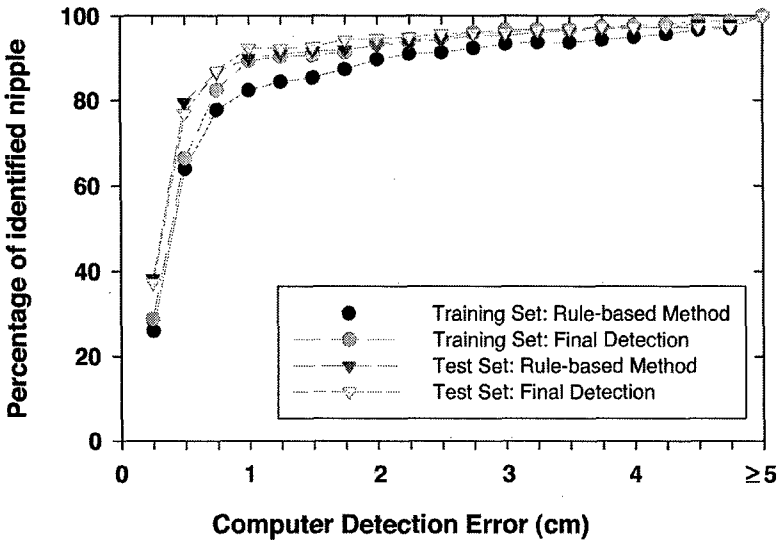


FIG. 9. The cumulative percentage of identified nipples with a computer detection error (Euclidean distance from the detected nipple location to the "gold standard") less than or equal to a certain distance for the visible nipple mammograms.

of which the identified nipple was within a certain distance from the gold standard for visible and invisible nipples, respectively. The computer performances at any detection error threshold can be obtained from these plots.

C. Observer variability for identifying invisible nipples

For the nipples that could not be positively identified, i.e., the invisible nipples, an estimated location was given by the radiologists based on visual assessment. The average of estimated nipple locations of two radiologists was used as the "gold standard" to reduce the subjective bias between radiologists. For the training set, if Radiologist 1's first reading, second reading, and the average of these two readings were compared to Radiologist 2's reading, the percentage of images with an agreement within 1 cm between the two estimated nipple locations was 84% (64/76), 79% (60/76), and 83% (63/76), respectively. If Radiologist 1's two readings were compared, the percentage of images with an agreement within 1 cm was 87% (66/76). However, if Radiologist 1's first reading, second reading, the average of these two readings, and Radiologist 2's reading were compared to the averaged "gold standard," the percentage of images with an agreement within 1 cm was 92% (70/76), 91% (69/76), 93% (71/76), and 99% (75/76), respectively. For the test set under the same conditions, the percentage of images with an agreement within 1 cm was 80% (55/69), 78% (54/69), and 78% (54/69) for the interobserver comparisons, 77% (53/69) for the intraobserver comparison, and 84% (58/69), 90% (62/69), 93% (64/69), and 96% (66/69) if the two radiologists' readings were compared to the averaged "gold standard." Figure 11 shows the histogram of intraobserver variation in marking the nipple locations by Radiologist 1 for the invisible nipple images in the training and test set.

IV. DISCUSSION

In this study, the resolution of the digitized mammograms was reduced to 800 μm $\times$ 800 μm for identification of the nipple locations both by the radiologists and by the computer. This low resolution was chosen in order to increase the computational efficiency. To verify that this resolution was sufficient for nipple identification, we performed a limited observer study to evaluate the dependence of the nipple visibility on pixel size. Eight full resolution images (50 μm $\times$ 50 μm) with invisible nipple (classified at 800 μm $\times$ 800 μm resolution) and one with a very subtle nipple profile were used. The images were subsampled to 100 μm, 200 μm, 400 μm, 600 μm, and 800 μm pixel size. One experienced MQSA radiologist who provided the gold standard described above was asked to visually inspect the nipple location on images of pixel size from 800 μm down to 100 μm individually. The windowing and zooming functions were used in the process of inspection. The observer study indicated that, if a nipple was not visible in a lower resolution image, for example, 800 μm, the nipple still could not be identified confidently by the radiologist on higher resolution images up to 100 μm. This result may be attributed to the fact that the size of the nipple is generally much larger than 800 μm $\times$ 800 μm. The visibility of the nipple is not limited by the resolution of the image at this pixel size. Most of the invisible nipples were caused by their nearly flat profiles, by the noise along the breast boundary, or by masking of the nipple behind dense tissue due to improper positioning. The smoothing with a box filter reduces the noise which may actually improve the visibility of objects that are not resolution-limited. The visibility of the nipples therefore was not improved by using higher resolution images.

The nipple identification method in this study assumes that most nipples are projected within 1 cm of the breast boundary on the mammogram. In our data set, based on radiologist's marking of nipple locations, we rejected 5 mammograms in which the nipple was located far away from the breast boundary due to skin folds or improperly positioned breast for imaging. The cases that contained a big breast exceeding the film area so that no nipple was projected on the mammograms were also rejected. Two experienced radiologists provided the gold standard by visually identifying

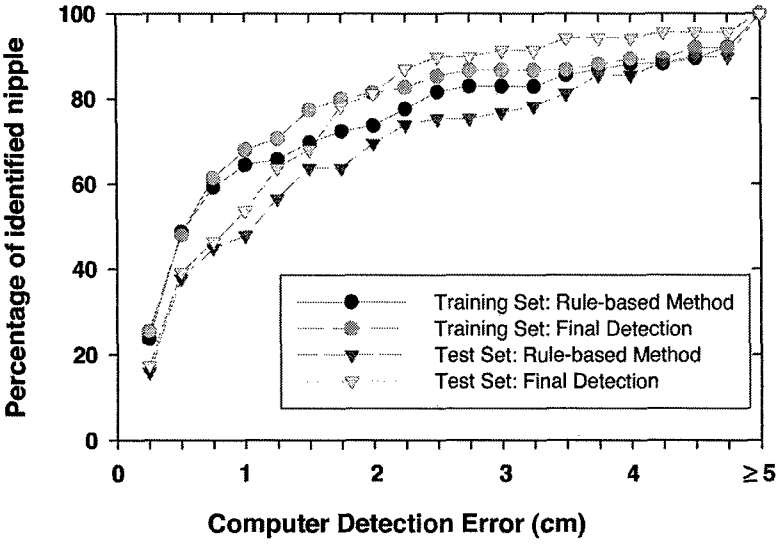


FIG. 10. The cumulative percentage of identified nipples with a computer detection error (Euclidean distance from the detected nipple location to the "gold standard") less than or equal to a certain distance for the invisible nipple mammograms.

the nipple location using a computer interface to display and adjust the contrast and brightness of the image. The nipple locations were marked by the radiologists at the center of the projected nipple image regardless of the size of the nipples which may vary from invisible to a diameter of larger than 1 cm. This means that the error of the computer detected nipple location from the gold standard mark would be larger for larger nipples because our computer method identified the nipple by searching along the breast boundary and the detected nipple location was marked at the breast boundary. For the nipples that could not be positively identified, the nipple location was given by radiologists' subjective visual estimation. From the comparison of inter- and intraobserver variability as described in the Results, it can be seen that Radiologist 2 had slightly higher agreement with the gold standard because most of the estimated nipple locations by Radiologist 2 were located between Radiologist 1's two readings. It can also be seen that the agreement between the two radiologists' readings and the "gold standard" was higher for the training set than that for the test set. This is in agreement with the performance achieved by our computer program in detecting nipple locations within 1 cm of the gold standard on the invisible nipple images, which was also higher for the training set (69.74%) than for the test set (53.62%). These results demonstrate that there were large variations in estimating the nipple locations for these difficult cases even by experienced radiologists.

The nipple detection method in this study depends primarily on nipple search along the breast boundary. At this stage, successful identification of the nipple depends on whether the breast boundary is tracked correctly. In the 744 mammograms used in our study, 110 nipples failed to be detected within 1 cm of the gold standard, of which 14.5% (16/110) of the boundary was rated as 1+ or 1−, and 1.8% (2/110) was rated as 2+ or 2−. In the 744 mammograms, the boundaries were rated as 1+ or 1− in 73 mammograms, 78.1% (57/73) of these nipples could be identified within 1 cm of the gold standard. For the 5 mammograms with worst boundary tracking (rated as 2+ or 2−), 60% (3/5) of the nipples still could be identified within 1 cm of the gold standard. Without using texture convergence analysis, 68.5% (50/73) and 60% (3/5) of these nipples were detected within 1 cm of the gold standard, respectively. It indicates that the nipple

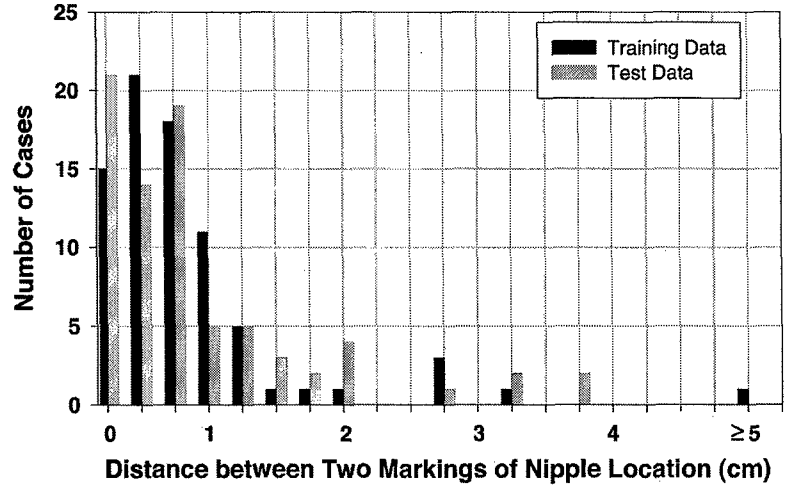


FIG. 11. Histogram of the intraobserver variations in marking the nipple locations by Radiologist 1 for the invisible nipple class.

detection in 7 of the images with boundary rated as 1+ or 1− failed at the stage of rule-based detection but was successful at the stage of texture convergence analysis for the mammograms. However, 2 mammograms with boundary rated as 2+ or 2− could not be correctly identified either by the rule-based method or by the texture convergence method.

There were large variations in the projected nipple images on the mammograms. For the nipples that were projected outside the breast boundary, the nipple should exhibit higher gray levels than the background pixels outside the boundary. For these convex nipples, the tracked breast boundary could depict a nipple shape. The shape depicted on the boundary was unique if no noise, such as fingerprint or artifacts on the film, affected the boundary tracking. In such cases,searching for the nipple shape along the boundary could find a reliable nipple location. However, some of the convex nipples had very poor signal-to-noise ratio due to scattered radiation. These mammograms often also had noisy boundary. Both factors could lead to false detection and thus large errors from the gold standard. For the situation when the nipple was projected inside the breast boundary, the detection was complicated by noise. Most of such noise was due to dense tissue structures near the breast boundary. Detecting the gray level changes along the breast boundary could potentially find the true nipple location. However, the false positives were higher in images of dense breasts with prominent structured noise.

For the cases that had low confidence in the detected nipple location by the rule-based method, the computer performed a texture convergence analysis based on the texture orientation of the dense glandular tissues or ducts near the nipple region. The texture feature analysis was found to be useful for improving the accuracy of nipple identification in this study. With our algorithm, 46.18% (139/301) of the visible nipples and 77.63% (59/76) of the invisible nipples in the training data set, and 72.73% (144/298) of the visible nipples and 89.86% (62/69) of the invisible nipples in the test data set could not be identified with high confidence by the rule-based method and the texture feature analysis was invoked. For these cases that had low confidence in the detection of nipple location by the rule-based method, 84.89% (118/139) of the visible nipples and 64.00% (38/59) of the invisible nipples in the training data set, and 85.42% (123/144) of the visible nipples and 55.00% (34/62) of the invisible nipples in the test data set could be identified within 1 cm of the gold standard by using the rule-based method in combination with texture convergence analysis. We applied a paired t-test to the detection errors on the subset of images for which the texture convergence analysis was used. The results indicated that the improvement in the accuracy was statistically significant for the visible nipple images in the training set ($p<0.002$) and the invisible nipple images in the test set ($p<0.005$), and did not achieve statistical significance for the visible nipple images in the test set ($p>0.87$) and the invisible nipple images in the training set ($p>0.68$).

In our study, the training and test sets were randomly selected from patient files. The results showed that, for the visible nipples, the algorithm performance achieved a higher accuracy (i.e., percentage of the detected nipples within 1 cm of the gold standard) in the test set (92.28%) than in the training set (89.37%). On the other hand, for the invisible nipples, the detection accuracy was higher in the training set (69.74%) than in the test set (53.62%). The different trends in the two nipple groups are most likely caused by sampling bias such that the visible nipple images in the test set were by chance somewhat easier to detect than those in the training set. To estimate the statistical significance of the difference in the algorithm performances between the training set and the test set, the bootstrap method was used to resample the training set 100 times and similarly for the test set. The mean and the standard deviation of the detection accuracy were then estimated from the bootstrap samples for the training set and for the test set. Using these estimated values, the unpaired t-test showed that the differences in the performance of our nipple detection method between the training set and the test set were statistically significant ($p<0.0001$) for both the visible and the invisible nipple groups. The estimated mean and standard deviation of the detection accuracy estimated from the resampled training and test sets and the corresponding p-values of the unpaired t-test are shown in Table II.

TABLE II. The unpaired t-test result which was used to estimate the statistical significance of the difference in the algorithm performances between the training set and the test set. The mean and standard deviation of the detection accuracy were estimated from the resample training and test data set using the bootstrap method.

		Mean	Standard deviation	p-value of unpaired t-test
Visible nipples	Training set	89.34%	1.87%	
	Test set	92.36%	1.58%	<0.0001
Invisible nipples	Training set	69.99%	5.06%	
	Test set	54.09%	6.09%	<0.0001

Although the performance of our nipple detection method is reasonable, further improvement in its accuracy is needed. One possible method may be first determining whether the breast contains very dense tissues, especially in the region posterior to the nipple, and weight the confidence of the texture convergence analysis accordingly. We will pursue this and other methods to improve the accuracy in future studies.

V. CONCLUSION

Accurate identification of nipple location on mammograms is challenging because of the variations in image quality and in the nipple projections, especially for the nipples that are nearly invisible on the mammograms. In this work, we developed a two-stage computerized nipple identification method to detect or estimate the nipple location. The results demonstrate that the visible nipples can be accurately detected by our computerized image analysis method. The nipple location can be reasonably estimated even if it is invisible. Automatic nipple identification will provide the foundation for multiple image analysis in CAD.

ACKNOWLEDGMENTS

This work was supported by USPHS Grant No. CA095153 and U. S. Army Medical Research and Material Command Grant No. DAMD17-02-1-0214. The content of this publication does not necessarily reflect the position of the funding agency, and no official endorsement of any equipment and product of any companies mentioned in this publication should be inferred.

[a]Author to whom correspondence should be addressed. Phone: 734-647-8552; Fax: 734-615-5513; electronic mail: chuan@umich.edu

[b]Current address: Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, Maryland 20857.

[1]S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics, 1998," Ca-Cancer J. Clin. **48**, 6–29 (1998).

[2]C. Byrne, C. R. Smart, C. Cherk, and W. H. Hartmann, "Survival advantage differences by age: Evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project," Cancer **74**, 301–310 (1994).

[3]H. C. Zuckerman, *The Role of Mammography in the Diagnosis of Breast Cancer*, in *Breast Cancer, Diagnosis and Treatment*, edited by I. M. Ariel and J. B. Cleary (McGraw–Hill, New York, 1987).

[4]S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of computerized mass detection on mammograms: Fusion of two-view information," Med. Phys. **29**, 238–247 (2002).

[5]B. Sahiner, H. P. Chan, L. M. Hadjiiski, M. A. Helvie, M. A. Roubidoux, and N. Petrick, "Computerized detection of microcalcifications on mammograms: Improved detection accuracy by combining features extracted from two mammographic views," Chicago, IL, November 30–December 5, 2003.

[6]L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, and M. A. Helvie, "Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis-local affine transformation for improved localization," Med. Phys. **28**, 1070–1079 (2001).

[7]R. Chandrasekhar and Y. Attikiouzel, "A simple method for automatically locating the nipple on mammograms," IEEE Trans. Med. Imaging **16**, 483–494 (1997).

[8]A. J. Mendez, P. G. Tahoces, M. J. Lado, M. Souto, J. L. Correa, and J. J. Vidal, "Automatic detection of breast border and nipple in digital mammograms," Comput. Methods Programs Biomed. **49**, 253–262 (1996).

[9]F. F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," Med. Phys. **21**, 445–452 (1994).

[10]A. R. Morton, H. P. Chan, and M. M. Goodsitt, "Automated model-guided breast segmentation algorithm," Med. Phys. **23**, 1107–1108 (1996).

[11]M. M. Goodsitt, H. P. Chan, B. Liu, A. R. Morton, S. V. Guru, S. Keshavmurthy, and N. Petrick, "Classification of compressed breast shape for the design of equalization filters in mammography," Med. Phys. **25**, 937–948 (1998).

[12]M. Worring and A. W. M. Smeulders, "Digital curvature estimation," CVGIP: Image Understand. **58**, 366–382 (1993).

[13]A. R. Rao and B. G. Schunck, "Computing oriented texture fields," CVGIP: Graph. Models Image Process. **53**, 157–185 (1991).

[14]A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," IEEE Trans. Image Process. **9**, 846–859 (2000).

[15]C. Zhou, H. P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, B. Sahiner, and L. M. Hadjiiski, "Computerized image analysis: Estimation of breast density on mammograms," Med. Phys. **28**, 1056–1069 (2001).

ROC study of the effect of stereoscopic imaging on assessment of breast lesions

Heang-Ping Chan,[a] Mitchell M. Goodsitt, Mark A. Helvie, Lubomir M. Hadjiiski, Justin T. Lydick, Marilyn A. Roubidoux, Janet E. Bailey, Alexis Nees, Caroline E. Blane, and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 12 November 2004; revised 15 January 2005; accepted for publication 22 January 2005; published 22 March 2005)

An observer performance study was conducted to evaluate the usefulness of assessing breast lesion characteristics with stereomammography. Stereoscopic image pairs of 158 breast biopsy tissue specimens were acquired with a GE Senographe 2000D full field digital mammography system using a 1.8× magnification geometry. A phantom-shift method equivalent to a stereo shift angle of ±3° relative to a central axis perpendicular to the detector was used. For each specimen, two pairs of stereo images were taken at approximately orthogonal orientations. The specimens contained either a mass, microcalcifications, both, or normal tissue. Based on pathological analysis, 39.9% of the specimens were found to contain malignancy. The digital specimen radiographs were displayed on a high resolution MegaScan CRT monitor driven by a DOME stereo display board using in-house developed software. Five MQSA radiologists participated as observers. Each observer read the 316 specimen stereo image pairs in a randomized order. For each case, the observer first read the monoscopic image and entered his/her confidence ratings on the presence of microcalcifications and/or masses, margin status, BI-RADS assessment, and the likelihood of malignancy. The corresponding stereoscopic images were then displayed on the same monitor and were viewed through stereoscopic LCD glasses. The observer was free to change the ratings in every category after stereoscopic reading. The ratings of the observers were analyzed by ROC methodology. For the 5 MQSA radiologists, the average A_z value for estimation of the likelihood of malignancy of the lesions improved from 0.70 for monoscopic reading to 0.72 ($p=0.04$) after stereoscopic reading, and the average A_z value for the presence of microcalcifications improved from 0.95 to 0.96 ($p=0.02$). The A_z value for the presence of masses improved from 0.80 to 0.82 after stereoscopic reading, but the difference fell short of statistical significance ($p=0.08$). The visual assessment of margin clearance was found to have very low correlation with microscopic analysis with or without stereoscopic reading. This study demonstrates the potential of using stereomammography to improve the detection and characterization of mammographic lesions. © *2005 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1870172]

Key words: digital mammography, stereoscopic imaging, specimen radiograph, observer performance study, ROC methodology

I. INTRODUCTION

Mammography is currently the only recommended imaging technique for breast cancer screening. However, mammographic sensitivity is often limited by the presence of dense breast parenchyma.[1] It has been reported that the false negative rate of mammography in dense breasts can be as high as 25%.[2,3] One of the main factors contributing to these missed cases is the camouflaging effect of the overlapping structures in the projection x-ray images. With the advent of high-resolution digital detectors for mammography, a number of new breast imaging techniques such as stereomammography,[4–12] digital tomosynthesis,[13–15] and computed tomography[16–18] are being developed in an effort to alleviate this problem. These techniques attempt to view the breast in three dimensions (3D) or to slice the breast volume into thin planes so as to reduce the superposition of breast tissue structures as imaged in two-dimensional (2D) projection mammograms. An observer performance study by Getty *et al.*[8] indicated that digital stereomammography improved the estimate of the probability of malignancy of mammographic lesions and allowed the detection of additional lesions that were obscured on screen-film mammograms. Rafferty *et al.*[19] also demonstrated that digital tomosynthesis mammograms could reveal additional lesions obscured by dense breast tissue and improved visualization of the margins and spiculations of masses.

Stereoscopic imaging requires acquisition of a left-eye image and a right-eye image. In conventional film-based stereoradiography, two film images were obtained by shifting the x-ray source, along a direction parallel to the image plane, to the left and the right of the central axis of the imaging system. When the two film images are placed properly and viewed so that the left eye sees only the left-eye film and the right eye sees only the right-eye film, the parallax between the two images creates the depth perception. Stereoscopic imaging was utilized for various types of radiographic

 0094-2405/2005/32(4)/1001/9/$22.50

examinations.[20–25] However, it did not receive widespread acceptance in clinical practice, mainly because of the doubled film cost and increased patient exposure.[26] In addition, radiologists had to read the stereoradiographs with a somewhat cumbersome film stereoscope or had to be trained to read the stereoradiographs without aid using a "cross-eyed" technique.

In recent years, direct digital detectors have become available for medical imaging. Stereoradiography may become a viable approach with digital imaging because there are no additional film costs. Furthermore, digital detectors have a linear response, wider dynamic range, and higher contrast sensitivity than screen-film systems so that good-quality digital stereo image pairs may be acquired at essentially the same total radiation dose as that for a conventional single-projection screen-film image. Maidment *et al.*[12] found that human eyes can integrate the noise in the left-eye and right-eye images such that the detectability of simulated low contrast objects on a uniform noisy background in a single image was comparable to that of viewing the left- and right-eye image pair when the total dose of the latter was about 1.1× of the dose of the single image. Maidment's experimental design evaluated the efficiency of noise reduction by binocular summation without utilizing the potential additional advantage of stereo depth perception in signal detection. It is likely that this additional advantage would further reduce the total dose requirements for stereo imaging to the same as or even lower than those for a single-projection image. Digital stereoscopic images can be viewed more conveniently than stereo film radiographs because of the electronic display. Different methods for displaying digital stereoscopic images are still being developed. One common method is to display the left-eye and right-eye images alternately at a very fast refresh rate on a monitor. The images are viewed with a pair of special goggles that typically consist of liquid crystal electronic shutters. The shutters are synchronized with the display so that the left eye of the reader is allowed to see only the left-eye image and the right eye is allowed to see only the right-eye image. For high-resolution medical images such as mammograms, no commercial stereo display systems are available at present.

Stereoradiography provides structural information of the object being viewed in 3D. It has been reported that the spatial distribution of microcalcifications may be associated with the malignant or benign nature of the cluster.[27,28] Masses may be better separated from the overlapping fibroglandular tissues in stereo than that in a 2D mammogram, making it easier to visualize the margin characteristics and determine whether spiculations are present. Therefore, stereomammography has the potential of providing additional diagnostic information that may improve the characterization of malignant and benign lesions and reduce unnecessary biopsies.

We are developing stereomammography techniques using a digital mammography system. In our previous studies, we examined the effects of stereo shift, geometric magnification, x-ray exposure, and display zooming on visual depth discrimination of crossing fibrils in stereo phantom images.[4,5,10] We found that a 2 mm depth discrimination could be achieved with over 90% accuracy on magnification images. We also investigated the accuracy of using a calibrated virtual cursor to measure the absolute depth of fibrils in stereoscopic images.[6,7,11] Our results showed that the average root-mean-square errors of depth measurements in stereo images with the virtual cursor ranged from 0.2 to 1.3 mm, depending on the stereo shift angle and the imaging geometry. These studies demonstrated that stereoscopic imaging can provide both qualitative depth discrimination and quantitative measurement of fibrous structures in a breast. In the present investigation, we conducted an observer performance study using receiver operating characteristic (ROC) methodology to investigate the effects of stereoscopic reading on the accuracy of detection and characterization of mammographic lesions using images of biopsied breast tissue specimens.

II. MATERIALS AND METHODS

A. Data set

Digital stereoscopic image pairs of the breast tissue specimens were acquired with a GE Senographe 2000D full field digital mammography (FFDM) system. The study was approved by the Institutional Review Board. The GE system uses a flat panel digital detector composed of a CsI:Tl scintillator and an amorphous-Si active matrix array. The detector has a pixel size of 100 μm × 100 μm and an output gray level resolution of 14 bits. The raw images are routinely processed with GE proprietary software and converted to 12 bit processed images. We employed a 1.8× magnification geometry (no grid, 0.15 mm focal spot) and a stereo shift angle of ±3° for imaging the stereoscopic specimen radiographs.

The conventional method for stereoradiography is to move the x-ray source to the left and the right of the central ray by a chosen stereo shift angle $\pm\theta^\circ$ (or stereo shift distance $\pm w$) for acquiring the left-eye and right-eye images. In the early days of radiography, it was determined by trial and error that a total tube shift equal to 10% of the focus-to-film distance produced satisfactory stereo results.[26] This is equivalent to a tube shift of about ±3° ($\cong \frac{1}{2}\tan^{-1}(0.1)$). In our previous studies,[4–7,9–11] we also found that ±3° would provide sufficient stereoscopic vision without causing excessive eye strain. The FFDM system was not designed for stereoscopic imaging. It does not have an electronic or mechanical lock mechanism to keep the x-ray tube stationary at the appropriate shift angle, nor do the collimator blades adjust to maintain complete coverage of the detector when the x-ray tube is shifted. We designed a stereo image acquisition method for phantoms and specimens in which the object is shifted instead of the focal spot. As illustrated in Fig. 1, the exposure geometry for the object relative to the focal spot when the focal spot is shifted to the left is equivalent to that when the focal spot is stationary and the object is shifted to the right by the same distance. Similarly, the geometry when the focal spot is shifted to the right is equivalent to that when the focal spot is stationary and the object is shifted to the left. A small error is caused by the slightly shorter focal-spot-to-detector distance in the object-shift geometry because the

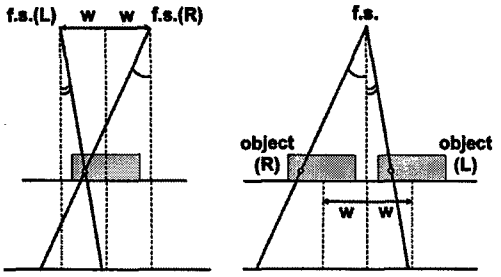


FIG. 1. Imaging geometry for acquisition of stereoscopic image pairs in magnification geometry. Left panel: a conventional "focal-spot shift" method in which the focal spot is shifted to the left and to the right of the central ray by a distance w to expose the left-eye and right-eye image. Right panel: an equivalent "object shift" method in which the object is shifted to the right and to the left of the central ray by the same distance, w. It can be seen that the image exposed by the f.s.(L) geometry is equivalent to that exposed by the object (L) geometry. Similarly, the image exposed by the f.s.(R) geometry is equivalent to that exposed by the object (R) geometry.

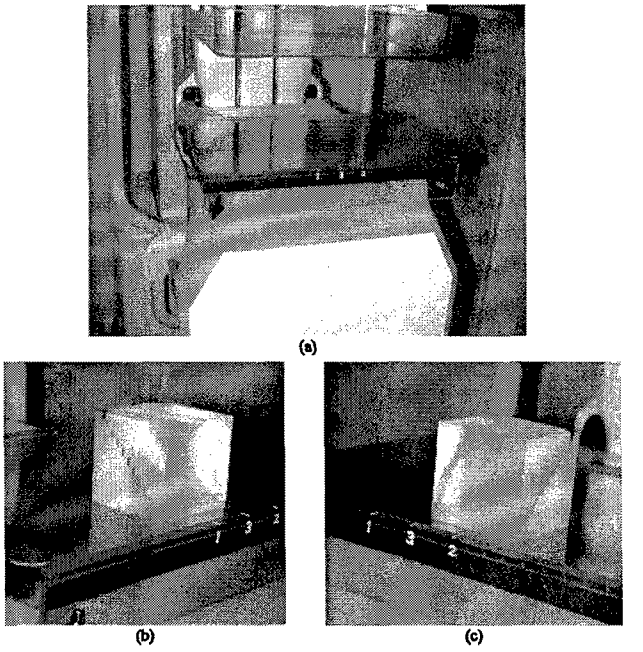


FIG. 2. The platform fits on the magnification stand of the FFDM system: (a) the sliding plate on top of the stationary base at the central position, marked as 3, (b) the sliding plate was shifted to the left position at 2.4 cm, marked as 1, and (c) the sliding plate was shifted to the right position at 2.4 cm, marked as 2. The stepwedge phantom shows where the tissue specimen would be placed.

x-ray focal spot moves along an arc. This error is estimated to be less than 0.1% for a ±3° stereo angle shift and a fulcrum of rotation at 46 cm from the focal spot. Using the geometry of the GE system and the ±3° stereo angle shift used in this study, the object shift distance, w, can be calculated to be ±2.4 cm from the central position. For a given stereo angle shift, the linear shift distance is the same for both the contact geometry and the magnification geometry. The phantom-shift technique was also used in our previous phantom studies.[10,11]

To facilitate the shifting of the object in a direction parallel to the chest wall (focal spot shift direction) for the FFDM system, we built a platform using Lexan plates shown in Fig. 2. The platform has a stationary base that fits on the magnification stand. The object is placed on a sliding plate on top of the base. The sliding plate can be moved manually between two guardrails in a direction parallel to the chest wall. The central position and the left and right shift distances were marked on the stationary base. The tissue specimens could therefore be moved to the desired left and right shift locations easily and precisely. Two fiducial markers (small metal rings) were affixed to the sliding plate. Their positions in the images were later used for alignment of the left-eye and right-eye images of the stereo pairs.

Consecutive biopsied breast tissue samples that were sent to the radiology department for specimen radiographs were imaged additionally with the stereoscopic technique if the FFDM system was available. The specimens were therefore random samples without selection. Each specimen could contain microcalcifications, mass, both, or normal tissue. Some specimens were obtained with ultrasound-guided biopsy of mammographically occult masses. The normal tissue was usually a result of a second biopsy to excise additional margins if the first tissue specimen was found to have a close margin. Two sets of stereo image pairs were acquired of each sample. These were acquired in approximately orthogonal orientations, whereby the second set was obtained by rolling the sample over by approximately 90°. The exposure techniques were manually chosen by mammography technologists. The mammography technologists were instructed to use high dose, identical techniques for the left-eye and right-eye images. The target/filter combinations were mainly Mo/Mo with Mo/Rh in some cases. The kilovoltage ranged from 24 to 27 kVp and the mAs ranged from 40 to 80 mAs, depending on the thickness of the tissue specimen.

All stereo image pairs were visually inspected for alignment and exposure by an experienced physicist. Some samples were rejected because of improper shift between the left-eye and right-eye images or improper exposure. All image pairs with good stereoscopic quality and exposure were included. This resulted in a total of 316 stereo image pairs from 158 specimens for the observer experiment. Based on pathological analysis 39.9% of the chosen samples were proven to contain malignancy. The lesion types and the number of lesions of each type for the samples used are listed in Table I. Examples of stereo image pairs of the tissue speci-

TABLE I. The lesion types and number of tissue specimens in each type.

Lesion	Malignant	Benign	Total
Mass	21	31	52
Microcalcifications	14	38	52
Both	24	9	33
No visible mass or microcalcifications	4	17	21
Total	63 (39.9%)	95 (60.1%)	158

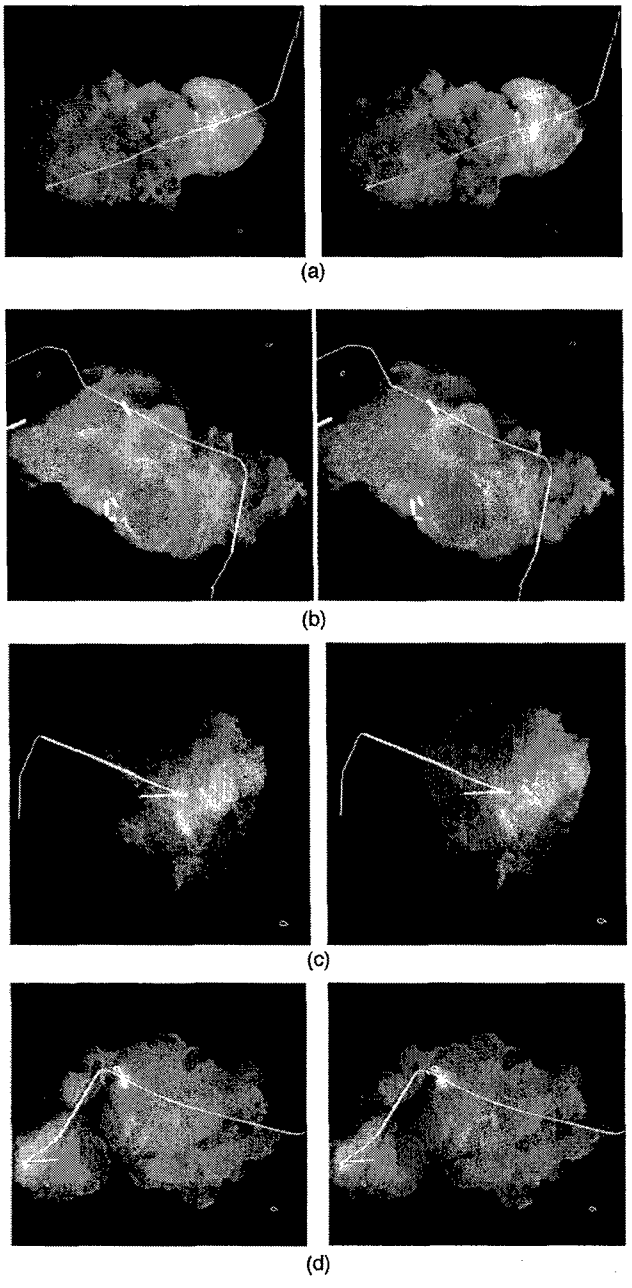


FIG. 3. Examples of stereo image pairs (left-eye image and right-eye image) of breast tissue specimens: (a) specimen with microcalcifications—invasive ductual carcinomas, (b) specimen with mass—invasive ductal carcinomas, (c) specimen with radial scar and microcalcifications—fibrocystic change, and (d) specimen with mass—fibrocystic change and fat necrosis.

mens are shown in Fig. 3.

B. Stereo image display

The images were displayed on a stereo workstation that consists of a MegaScan 8 mega-pixel CRT monitor driven by a Dome Md8-4820-LS stereoscopic board and a PC. The monitor was adjusted with a photometer to meet the DICOM grayscale standards, and the room lights were dimmed to a very low level during the observer studies. The system can display full-field (2300×1800 pixels) digital mammograms at a refresh rate of 120 Hz. It operates in a page flipping stereoscopic mode with the left- and right-eye images displayed alternately. A pair of CrystalEyes LCD stereoscopic glasses was used for viewing the stereoscopic images. The stereo images were displayed with in-house developed software that provided functions to shift and align the left-eye and right-eye images, adjust the contrast and brightness, and store the selected alignment and windowing settings. The stereo display workstation is shown in Fig. 4. The physicist selected and saved the settings for each image pair which became the default settings when this image pair was displayed the next time. The same display conditions could therefore be used for all radiologists in the observer study. The radiologist had the option of adjusting the window settings if they deemed it necessary. The software could also switch the display to show the left-eye image alone or the right-eye image alone so that the observer could read the monoscopic image and the stereoscopic images of the same case sequentially, as designed for the observer experiment described in the following.

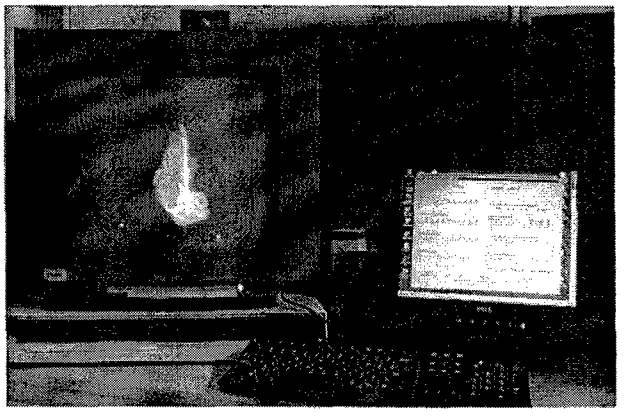


FIG. 4. Stereo display workstation composed of a MegaScan 8 mega-pixel monitor driven by a Dome Md8-4820-LS stereoscopic board and a PC. The system can display full-field (2300×1800 pixels) digital mammograms at a refresh rate of 120 Hz.

C. Observer performance study

A user interface was designed for the observer experiment. The user interface displayed images sequentially according to an input list. Slide bars were provided to record the observer's confidence ratings (scale of 1–100) regarding the presence of a mass, the presence of calcifications, the likelihood of malignancy of the lesion if present, the likelihood of the margin being clear. The observers were also asked to provide an assessment of malignancy in terms of the BI-RADS categories (1=negative, 2=benign, 3=probably benign, 4=suspicious, 5=highly suggestive of malignancy),[29] and a visual estimate of the margin clearance (0=positive margin, 1=0–2 mm, 2=2–5 mm, 3=greater than 5 mm). Five Mammography Quality Standards Act (MQSA) qualified radiologists participated in the experiment. The experiment was designed to have each observer

read the 316 specimen images in two sessions. The two views of each specimen were read independently and were arranged to be read in the two separate sessions to reduce the possibility of memorization. It may be noted that this was not equivalent to using 316 truly independent samples in the observer experiment. This increased the sample size but the possible correlation between the two views may cause a slight underestimation of the variances in the data. The reading sequence was systematically arranged in a counterbalanced design so that no specific cases were read by all observers always in the first or the second session. The case reading order was different for each observer. The observers first read the left-eye image alone as a monoscopic image and entered their assessments in all categories. The stereoscopic images were then displayed and were read with the LCD glasses. The observers were free to change their ratings in every category after reading the stereoscopic images. The observers were allowed unlimited time to read each case. They were also free to break the reading sessions into shorter ones. The radiologists were informed of the fact that the samples were randomly collected from the biopsied tissue specimens so that the proportion of malignant and benign cases would be similar to that in their clinical practice. They were therefore also aware that some specimens could be found to be negative for lesions or malignancy by pathological analysis.

Before a radiologist was recruited as observer, he/she underwent a standard Randot Circles Stereo test (Stereo Optical Co., Inc., Chicago, IL) to evaluate their stereo acuity. The reader viewed ten sets of circles on the test pattern through polarized glasses. Each set contained three circles, one of which would appear to be at a different depth from the others when viewed stereoscopically. The reader was asked to identify the circle that stood out in each of the ten sets. All radiologists participated in our observer performance study could correctly identify 9 to 10 of the circles, indicating that their level of stereopsis was at least 30 s of arc at a viewing distance of 16 in. Prior to reading the test stereo images, the observer also participated in a training session to become familiar with the reading task and the user interface.

D. Data analysis

The confidence ratings and the BI-RADS assessments of the observers were analyzed with the LABMRMC program.[30] The area under the ROC curve, A_z, and the partial area index above a sensitivity of 0.90, $A_z^{(0.9)}$, were used to compare the performance between monoscopic reading and monoscopic assisted with stereoscopic reading. The statistical significance of the difference in A_z between the two was estimated by the two-tailed p-value from the LABMRMC program and the Student's paired t-test. The average A_z and $A_z^{(0.9)}$ values were obtained from the average ROC curve that was derived from the average slope and intercept parameters of the individual readers' ROC curves. For the classification of malignant and benign lesions, all samples were analyzed together regardless of the lesion type.

TABLE II. Performance of radiologists in detecting microcalcifications in the tissue specimens with monoscopic (single projection) reading and with additional stereoscopic reading. The average A_z and $A_z^{(0.9)}$ were derived from the average a and b parameters of the individual ROC curves. The improvements in A_z and $A_z^{(0.9)}$ were both statistically significant with $p=0.02$ and $p=0.004$, respectively.

	A_z		$A_z^{(0.9)}$	
Radiologist	Monoscopic	With stereo	Monoscopic	With stereo
1	0.97±0.01	0.98±0.01	0.76	0.79
2	0.95±0.01	0.96±0.01	0.58	0.67
3	0.94±0.02	0.95±0.02	0.48	0.55
4	0.94±0.02	0.95±0.01	0.58	0.65
5	0.92±0.02	0.92±0.02	0.30	0.36
Average	0.95	0.96	0.57	0.63

For the analysis of the visual assessment of the margin status of the specimens in comparison with pathologists' analysis, we first combined the margin assessments from the two orthogonal views of the same specimen by taking the minimum margin clearance seen by the radiologist in the two views. This simulated the situation in which the radiologist was allowed to see the margins from the two different projections and estimated the minimum margin clearance from all visible borders, as they do in reading specimen radiographs in their routine clinical practice. The correlation of the radiologists' assessment of margin clearance with the result of pathological analysis was evaluated by the Pearson's correlation coefficient. Since pathological reports included margin assessment only for malignant lesions, only this subset of cases was used in the correlation analysis.

III. RESULTS

The radiologists' accuracy in detection of microcalcifications in the specimen by reading a single-projection image in comparison to that with additional stereoscopic reading is shown in Table II. In this ROC analysis, all samples with microcalcifications (malignant and benign) were considered to be positive cases. The samples with mass alone or without either mass or microcalcifications were treated as negative cases with respect to microcalcifications. The detection of microcalcifications in the small volume of tissue specimens appeared to be easy with or without stereoscopic reading. The A_z values for the five radiologists ranged from 0.92 to 0.97 with an average of 0.95 for monoscopic reading. Nevertheless, the radiologists still improved their performance with additional stereoscopic reading, with the A_z values ranging from 0.92 to 0.98 and an average of 0.96. The improvement, although modest, was consistent over all radiologists (the A_z value of Radiologist 5 improved from 0.918 to 0.922). The partial area index $A_z^{(0.9)}$ values for the radiologists were also high, ranging from 0.30 to 0.76 with monoscopic reading and improved to a range of 0.36 to 0.79 with

TABLE III. Performance of radiologists in detecting masses in the tissue specimens with monoscopic (single projection) reading and with additional stereoscopic reading. The average A_z and $A_z^{(0.9)}$ were derived from the average a and b parameters of the individual ROC curves. The improvements in A_z and $A_z^{(0.9)}$ both fell short of statistical significance with $p=0.08$ and $p=0.11$, respectively.

	A_z		$A_z^{(0.9)}$	
Radiologist	Monoscopic	With stereo	Monoscopic	With stereo
1	0.83±0.02	0.84±0.02	0.19	0.22
2	0.75±0.03	0.79±0.02	0.11	0.18
3	0.81±0.02	0.82±0.02	0.24	0.28
4	0.83±0.03	0.83±0.03	0.25	0.24
5	0.80±0.03	0.81±0.02	0.16	0.17
Average	0.80	0.82	0.19	0.22

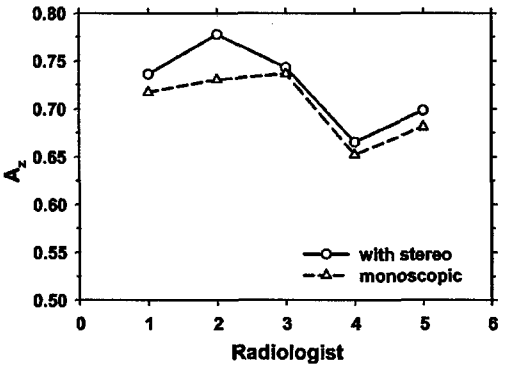


FIG. 5. The area under the ROC curves for the five radiologists for classification of malignant and benign lesions. The observers show a modest but consistent improvement in performance with additional stereoscopic viewing.

additional stereoscopic reading. The improvements in A_z and $A_z^{(0.9)}$ were both statistically significant with $p=0.02$ for A_z and $p=0.004$ for $A_z^{(0.9)}$.

The radiologists' accuracy in detection of masses with the two reading conditions is compared in Table III. Similar to the ROC analysis for microcalcifications, all samples with masses were considered positive. The samples with microcalcifications alone or without either mass or microcalcifications were considered negative for masses. For monoscopic reading, the A_z values of the radiologists ranged from 0.75 to 0.83 with an average A_z of 0.80. With additional stereoscopic reading, the A_z values for four of the five radiologists improved. The A_z ranged from 0.79 to 0.84 and the average A_z was improved to 0.82. However, the improvements in both A_z and $A_z^{(0.9)}$ fell short of statistical significance with $p=0.08$ and $p=0.11$, respectively.

Table IV shows the comparison of the radiologists' assessments of the likelihood of malignancy of the tissue specimens with and without stereoscopic reading. With monoscopic reading, the A_z values of the radiologists ranged from 0.65 to 0.74. Their accuracy improved significantly ($p=0.04$) with additional stereoscopic reading to the range of 0.67 to 0.78 (Fig. 5). The partial area index $A_z^{(0.9)}$ also improved significantly ($p=0.04$) from a range of 0.07 to 0.12 to a range of 0.09 to 0.19.

Table V shows the changes in BI-RADS categories with stereoscopic reading. Since the BI-RADS assessment of categories 3 or above indicates the need of call-back for further evaluation, and categories 4 and 5 indicate a recommendation for biopsy, we summarized the changes in the BI-RADS categories across the threshold between categories 1, 2 and 3, 4, 5, and the threshold between categories 1, 2, 3 and 4, 5. By counting the number of lesions having an increase in the categories across the threshold as positive and a decrease as negative, the average number of lesions that had significant changes in BI-RADS categories over the five radiologists for the malignant lesions and the benign lesions were calculated. The results revealed that the radiologists improved their assessments of malignant lesions with stereoscopic reading. For malignant lesions, the BI-RADIS assessments for an average of 1.6 lesions (1.6/63=2.5%) per radiologist were changed from negative or benign to call-back, and an average of 2.2 lesions (2.2/63=3.5%) per radiologist were changed from categories 1, 2, and 3 to recommendation for biopsy. However, for benign lesions there were also increases in call-back and biopsy recommendations but the

TABLE IV. Performance of radiologists in classification of malignant and benign lesions in the tissue specimens with monoscopic (single projection) reading and with additional stereoscopic reading. The average A_z and $A_z^{(0.9)}$ were derived from the average a and b parameters of the individual ROC curves. The improvements in A_z and $A_z^{(0.9)}$ were both statistically significant with $p=0.04$ and $p=0.04$, respectively.

	A_z		$A_z^{(0.9)}$	
Radiologist	Monoscopic	With stereo	Monoscopic	With stereo
1	0.72±0.03	0.74±0.03	0.07	0.09
2	0.73±0.03	0.78±0.03	0.12	0.19
3	0.74±0.03	0.74±0.03	0.09	0.11
4	0.65±0.03	0.67±0.03	0.10	0.11
5	0.68±0.03	0.70±0.03	0.10	0.13
Average	0.70	0.72	0.10	0.13

TABLE V. The average number of lesions per radiologist of which the BI-RADS category was changed after stereoscopic reading. BI-RADS categories 3 or above represent a call-back and categories 4 or above represent biopsy recommendation. Positive change indicated an increase in the number of lesions from the lower to the higher categories and negative change indicated a decrease.

	Average number of lesions per radiologist	
Change in BI-RADS assessment	From categories 1, 2 to categories 3, 4, 5	From categories 1, 2, 3 to categories 4, 5
Malignant lesions	1.6	2.2
Benign lesions	1.2	0.4

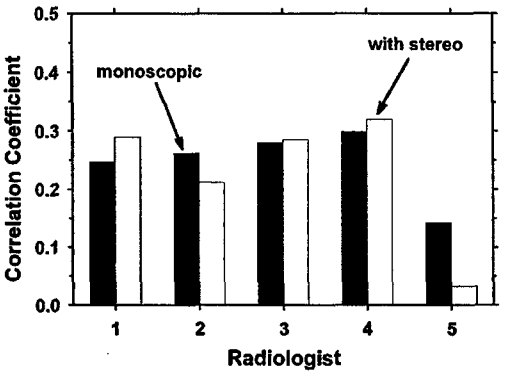


FIG. 6. The correlation coefficients between the radiologists' assessment of margin clearance and pathological analysis. The black bars were obtained with monoscopic reading, the white bars were obtained with additional stereoscopic reading.

changes were less, with an average of 1.2 (1.2/95=1.3%) and 0.4 (0.4/95=0.4%) lesions per radiologist for the two types of changes, respectively.

The correlation coefficients of the radiologists' assessment of margin clearance with pathological analysis are plotted in Fig. 6. The assessment of margin status visually in tissue specimens was found to be very unreliable. The correlation coefficients for all radiologists with or without stereoscopic reading were below about 0.3.

IV. DISCUSSION

The purpose of our study was to evaluate the potential advantages of stereo full-field digital mammography for the detection and characterization of breast lesions. Because of the difficulty of collecting a large data set of stereoscopic whole breast mammograms with lesions, we used stereo specimen radiographs for this preliminary study. Detection and characterization of lesions on specimen radiographs is different from similar tasks on FFDMs. Since the location of the lesion is confined to a smaller and thinner tissue sample than the whole breast, specimen radiographs should already provide superior visibility of lesion characteristics as compared to whole breast mammograms. Therefore, these are more difficult tasks for achieving improvements in the detection and characterization of the lesions. Nevertheless, our results indicate that the additional stereoscopic reading did improve the visualization of lesions and the accuracy of assessing their malignant or benign characteristics in specimen radiographs. Although the results cannot be generalized directly to reading whole breast mammograms, the potential for information gain and improvement in accuracy with stereoscopic reading have been demonstrated. In a study by Getty *et al.*[8] comparing the characterization of mammographic lesions on film mammograms alone to that with additional reading of whole breast stereomammograms, they observed an improvement in A_z from 0.83 to 0.86. Their slightly larger improvement in A_z with whole breast mammograms than that obtained in our study appears to corroborate our expectations.

The observer performance results indicated that the data set used in this study was quite difficult even in specimen radiographs. The mass detection task was challenging even for experienced radiologists, probably because the samples contained a variety of abnormal and normal specimens including focal densities and mammographically occult masses that were imaged with ultrasound during the wire localization procedure. The characterization of malignant and benign lesions was also difficult because these lesions had been recommended for biopsy so that they all appeared to be suspicious to some degree. There were also cases in which the lesions were partially removed by core biopsy so that the appearance might not be typical. The variety of cases was included because the ROC experiment measured the relative improvement with additional stereoscopic reading for the given set of samples rather than the absolute performance of the radiologists in clinical practice.

Breast tissue specimens are routinely radiographed and read by radiologists to determine primarily if the lesion recommended for biopsy is included in the specimen and secondarily if the cancer extends to the margin in lumpectomy cases. The results are used by the surgeon to determine if additional excision is needed. The low correlation between the visual assessment of margin clearance with the pathologists' report is somewhat unexpected. It therefore indicates that visual assessment of margin status does not correspond very well with microscopic analysis. It is likely that the specimen radiograph is useful for estimating whether the lesion is far from the specimen's boundaries. However, if the lesion is close to the margin, i.e., within a few millimeters, specimen radiographs are not capable of showing whether microscopic amounts of malignant tissues are present at the boundary.

In this study we used a sequential reading method, namely, the observer first read with monoscopic viewing and provided their ratings, and this was immediately followed with stereoscopic viewing and second ratings. The second ratings therefore represented diagnostic decisions resulting from a combination of the information from the conventional monoscopic reading with that from the additional stereoscopic reading. This will likely be the reading mode used if stereoradiographs are available clinically because the left-eye and right-eye images are readily available for monoscopic viewing and because there is no need to trade off any existing benefits of conventional reading in exchange for the stereoscopic viewing. The radiologists may switch between the monoscopic and the stereoscopic images to extract complementary information or to confirm their observations. This information gain may be obtained without or with a minimal increase in patient exposure compared with current screen-film mammographic techniques. Further studies of interest include comparisons of the detection and characterization of lesions under the following sequential reading conditions: (1) monoscopic reading of either the left-eye or the right-eye image alone, (2) monoscopic readings of both the left-eye and right-eye images by switching back and forth between the two, and (3) with additional stereoscopic reading of the image pair. These comparisons will reveal if the slight shift

in the perspective obtained from monoscopic readings of both the left-eye and right-eye images will in itself provide sufficient information to improve the detection and characterization performances or if the additional stereoscopic reading with depth perception is essential. Another study of interest is a comparison of monoscopic readings of the two orthogonal views of the specimens with stereoscopic reading of one of the views or both views. This study will reveal if the 3D information obtained from orthogonal views is superior to that from stereoscopic reading of one of the views or if additional information can still be gained from stereoscopic reading of both views. Likewise, a comparison of monoscopic readings of CC view and MLO view mammograms to stereoscopic reading of the MLO view mammogram alone or both views will be an interesting study to evaluate how stereomammography may be implemented in clinical practice.

To simplify image acquisition and the observer experiment, we used the left-eye image of the stereo pair as the monoscopic image for reading. Since the stereo shift angle is only ±3°, the difference in projection between an image taken at the central position (no-shift) and the left-eye (or the right-eye) image is very small. Each image of the stereo pair should be very similar to the central image. Furthermore, we instructed the technologists to use exposure techniques much higher than those used for a conventional specimen radiograph. The use of high dose techniques was intended to obtain monoscopic images of which the image quality would not be limited by quantum noise. This experimental design reduces the likelihood that the information gain with stereoscopic reading is due to the reduced noise when two monoscopic images were integrated into the stereoscopic image. Although it is difficult to perform a quantitative measurement to prove that this was indeed the case, all monoscopic images were visually evaluated and only low noise, high quality images were accepted as case samples for the observer experiment.

We adjusted the display monitor with a photometer to meet the DICOM grayscale standards. We did not attempt to take into account the attenuation by the LCD glasses in the adjustment because there are no DICOM standards for setting up a stereo display at present. The LCD glasses do degrade the perceived image quality to some extent, such as a reduction in brightness and an increase in noise. However, since the degradation would have a negative impact on stereoscopic reading, one may expect that the advantages of stereoscopic reading would be even greater than those observed in our study if the degradation could be compensated for or if better stereoscopic viewing methods (e.g., higher transmission stereo glasses or no glasses) become available in the future.

One of the expectations for developing 3D imaging techniques such as stereomammography for screening is to reduce recalls. In conventional mammography, many recalls are caused by superimposition of dense tissue mimicking masses and insignificant calcifications appearing to be clustered due to a lack of 3D spatial distribution information. In our study, analysis of the BI-RADS assessments indicates that the increase in detection sensitivity is accompanied by a slight increase in recalls. It is not known how the reading of specimen radiographs in a laboratory experiment would translate to clinical applications. However, the observed improvements in the ROC curves indicate that there were true improvements in the performances of the radiologists with additional stereoscopic reading and that the radiologists did not simply relax the decision thresholds along their original ROC curves, which would also result in an increase in sensitivity and a decrease in specificity. The improvements in the ROC curves show the promise that, if the radiologists become more experienced with stereomammography and more confident in utilizing the additional 3D information for assessing the lesions, they may be able to adjust their decision thresholds along the resulting higher ROC curves such that the sensitivity will be gained without a tradeoff, or even with an increase, in specificity in comparison to their decisions along the lower ROC curves associated with monoscopic reading alone. Further studies will be needed to investigate if this can be realized and thus lead to a reduction in recalls.

One limitation of stereoscopic viewing is that human eyes vary in their stereo acuity, although it is believed that stereo acuity may improve with training. The radiologists participated in this study were impressed by the 3D appearance of the stereoscopic images. The image quality of our stereo display workstation is excellent without perceivable flicker. However, some of the radiologists still experienced eye fatigue if the reading time was long. These problems may be alleviated with a different display method or viewing electronics as well as improved reader ergonomic factors.

V. CONCLUSION

We have performed an observer performance study using ROC methodology to evaluate the improvement in mammographic lesion detection and characterization by stereoscopic reading. Our results indicated that statistically significant (two-tailed $p < 0.05$) improvements were achieved for detection of microcalcifications and for classification of malignant and benign lesions. The detection of masses was also improved but the improvement fell short of statistical significance. This study demonstrates the potential of using stereomammography to improve the detection and characterization of mammographic lesions.

ACKNOWLEDGMENTS

This work was supported by U.S. Army Medical Research and Materiel Command Grant Nos. DAMD17-98-1-8210, DAMD17-02-1-0214, and DAMD17-99-1-9294. The content of this publication does not necessarily reflect the position of the funding agency, and no official endorsement of any equipment and product of any companies mentioned in this publication should be inferred. The authors are grateful to Dr. Charles E. Metz for the LABMRMC program.

[a]Electronic mail: chanhp@umich.edu

[1]V. P. Jackson, R. E. Hendrick, S. A. Feig, and D. B. Kopans, "Imaging of

the radiographically dense breast," Radiology **188**, 297–301 (1993).
[2]R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," Radiology **184**, 613–617 (1992).
[3]M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," Clin. Radiol. **44**, 13–15 (1991).
[4]H. P. Chan, M. M. Goodsitt, J. M. Sullivan, K. L. Darner, and L. M. Hadjiiski, "Depth perception in digital stereoscopic mammography," Atlanta, GA, 8–12 June, 2000.
[5]H. P. Chan, M. M. Goodsitt, K. L. Darner, J. M. Sullivan, L. M. Hadjiiski, N. Petrick, and B. Sahiner, "Effects of stereoscopic imaging technique on depth discrimination," *The Fifth International Workshop on Digital Mammography*, Toronto, Canada, 11–14 June, 2000 (Medical Physics, Madison, WI), pp. 13–18.
[6]M. M. Goodsitt, H. P. Chan, and L. M. Hadjiiski, "Stereomammography: Evaluation of depth perception using a virtual 3D cursor," Med. Phys. **27**, 1305–1310 (2000).
[7]M. M. Goodsitt, H. P. Chan, J. M. Sullivan, K. L. Darner, and L. M. Hadjiiski, "Evaluation of the effect of virtual cursor shape on depth measurements in digital stereomammograms," *The Fifth International Workshop on Digital Mammography*, Toronto, Canada, 11–14 June, 2000 (Medical Physics, Madison, WI), pp. 45–50.
[8]D. J. Getty, R. M. Pickett, and C. J. D'Orsi, *Stereoscopic Digital Mammography: Improving Detection and Diagnosis of Breast Cancer*, Berlin, 27–30 June, 2001, International Congress Series Vol. 1230 (Elsevier, Amsterdam), pp. 506–511.
[9]H. P. Chan, M. M. Goodsitt, L. Hadjiiski, M. A. Roubidoux, J. E. Bailey, M. A. Helvie, J. T. Lydick, and B. Sahiner, "ROC study comparing radiologists' performances in evaluating breast lesions on stereoscopic and single-projection digital specimen mammograms," Med. Phys. **30**, 1456 (abstract) (2003).
[10]H. P. Chan, M. M. Goodsitt, L. M. Hadjiiski, J. E. Bailey, K. Klein, K. L. Darner, and B. Sahiner, "Effects of magnification and zooming on depth perception in digital stereomammography: An observer performance study," Phys. Med. Biol. **48**, 3721–3734 (2003).
[11]M. M. Goodsitt, H. P. Chan, K. L. Darner, and L. M. Hadjiiski, "The effects of stereo shift angle, geometric magnification and display zoom on depth measurements in digital stereomammography," Med. Phys. **29**, 2725–2734 (2002).
[12]A. D. A. Maidment, P. R. Bakic, and M. Albert, "Effects of quantum noise and binocular summation on dose requirements in stereoradiography," Med. Phys. **30**, 3061–3071 (2003).
[13]L. T. Niklason, B. T. Christian, L. E. Niklason, D. B. Kopans, D. E. Castleberry, B. H. Opsahl-Ong, C. E. Landberg, P. J. Slanetz *et al.*, "Digital tomosynthesis in breast imaging," Radiology **205**, 399–406 (1997).
[14]R. L. Webber, H. R. Underhill, and R. I. Freimanis, "A controlled evaluation of tuned-aperture computed tomography applied to digital spot mammography," J. Digit Imaging **13**, 90–97 (2000).
[15]S. Suryanarayanan, A. Karellas, S. Vedantham, S. P. Baker, S. J. Glick, C. J. D'Orsi, and R. L. Webber, "Evaluation of linear and nonlinear tomosynthetic reconstruction methods in digital mammography," Acad. Radiol. **8**, 219–224 (2001).
[16]V. Raptopoulos, J. K. Baum, M. Hochman, A. Karellas, M. J. Houlihan, and C. J. D'Orsi, "High resolution CT mammography of surgical biopsy specimens," J. Comput. Assist. Tomogr. **20**, 179–184 (1996).
[17]J. M. Boone, T. R. Nelson, and J. A. Seibert, "The potential for breast CT," Med. Phys. **28**, 1246 (abstract) (2001).
[18]J. M. Boone, T. R. Nelson, K. K. Lindfors, and J. A. Seibert, "Dedicated breast CT: Radiation dose and image quality evaluation," Radiology **221**, 657–667 (2001).
[19]E. A. Rafferty, D. Georgian-Smith, D. B. Kopans, D. A. Hall, R. Moore, and T. Wu, "Comparison of full-field digital tomosynthesis with two view conventional film screen mammography in the prediction of lesion malignancy," Radiology **225(P)**, 268 (2002).
[20]K. Doi, N. J. Patronas, E. E. Duda, E. Geldner, and K. Dietz, "X-ray imaging of blood vessels to the brain by use of magnification stereoscopic technique," Adv. Neurol. **30**, 175–189 (1981).
[21]C. A. Kelsey, R. D. Moseley, S. A. Mettler, and D. E. Briscoe, "Cost-effectiveness of stereoscopic radiographs in detection of lung nodules," Radiology **142**, 611–613 (1982).
[22]K. Doi and E. E. Duda, "Detectability of depth information by use of magnification stereoscopic technique in cerebral angiography," Radiology **146**, 91–95 (1983).
[23]Y. Higashida, Y. Hirata, R. Saito, S. Doudanuki, H. Bussaka, and M. Takahashi, "Depth determination on stereoscopic digital subtraction angiograms," Radiology **168**, 560–562 (1988).
[24]M. C. Trocme, A. H. Sather, and K. N. An, "A biplanar cephalometric stereoradiography technique," Am. J. Orthod. Dentofacial Orthop. **98**, 168–175 (1990).
[25]J. I. Ragnarsson and J. Karrholm, "Factors influencing postoperative movement in displaced femoral neck fractures: evaluation by conventional radiography and stereoradiography," J. Orthop. Trauma **6**, 152–158 (1992).
[26]T. S. Curry, J. E. Dowdey, and R. C. Murry, *Christensen's Physics of Diagnostic Radiology*, 4th ed. (Lea & Febiger, Philadelphia, PA, 1992).
[27]A. D. A. Maidment, M. Albert, E. F. Conant, C. W. Piccoli, and P. A. McCue, "Prototype workstation for 3-D diagnosis of breast calcifications," Radiology **201(P)**, 556 (1996).
[28]E. F. Conant, A. D. Maidment, M. Albert, C. W. Piccoli, S. A. Nussbaum, and P. A. McCue, "Small field-of-view digital imaging of breast calcifications: method to improve diagnostic specificity," Radiology **201(P)**, 369 (1996).
[29]*American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)* (American College of Radiology, Reston, VA, 2003).
[30]D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," Radiology **27**, 723–731 (1992).

Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis

Jun Wei[a], Berkman Sahiner[a], Lubomir M. Hadjiiski[a], Heang-Ping Chan[a],
Nicholas Petrick[b], Mark A. Helvie[a], Chuan Zhou[a], Zhanyu Ge[a]
[a]Department of Radiology, University of Michigan, Ann Arbor
[b]Center of Devices and Radiological Health, U. S. Food and Drug Administration, Rockville, MD

ABSTRACT

Several full-field digital mammography (FFDM) systems have been approved for clinical applications. It is important to develop a CAD system that can easily be adapted to images acquired by FFDM systems from different manufacturers. To develop a CAD system that is independent of the FFDM manufacturer's proprietary preprocessing methods, we used the raw FFDM image as input and developed a multi-resolution preprocessing scheme for image enhancement. Our CAD system performed prescreening to identify mass candidates, segmented the suspicious structures, extracted morphological and texture features, and then classified masses and normal tissue. In this study, we investigated the use of a two-stage gradient field analysis to identify suspicious masses, and the effectiveness of a new gradient field feature extracted from each suspicious object for false positive (FP) reduction. A data set of 104 cases with 243 images acquired with a GE FFDM system was collected. Most cases had two mammographic views, except for 12 cases that had three views and 1 case with only one view. The data set contained 106 masses. The true locations of the masses were identified by an experienced radiologist. Using free-response receiver operating characteristic (FROC) analysis, it was found that our CAD system achieved a cased-based sensitivity of 70%, 80%, and 88% at 0.8, 1.3, and 1.7 FP marks/image, respectively. The high performance indicated the usefulness of the new gradient field analysis method.

Keywords: Computer-aided diagnosis (CAD), Full field digital mammography (FFDM), Gradient field analysis

1. INTRODUCTION

Breast cancer is one of the leading causes of death among American women between 40 to 55 years of age[1 2-4]. It has been reported that early diagnosis and treatment significantly can improve the chance of survival for patients with breast cancer[3-6]. Although mammography is the best available screening tool for detection of breast cancers, studies indicate that a substantial fraction of breast cancers that are visible upon retrospective analyses of the images are not detected initially[7-12]. Computer-aided diagnosis (CAD) is considered to be one of the promising approaches that may improve the sensitivity of mammography [13,14]. Computer-aided lesion detection can be used during screening to reduce oversight of suspicious lesions that warrant further work-up. It has been shown that CAD can improve radiologists' detection accuracy significantly [15-17].

Most of mammographic CAD algorithms developed so far are based on digitized mammograms. In the last few years, full-field digital mammography (FFDM) technology has advanced rapidly because of the potential of digital imaging to improve breast cancer detection. Several FFDM systems have become commercially available. We have developed a CAD system for the detection of masses on digitized mammograms in our previous study[18,19]. We are developing a mass detection system for mammograms acquired directly by an FFDM system. In this study, we are investigating the use of gradient field analysis to improve the performance of our mass detection system for FFDMs.

Medical Imaging 2004: Image Processing, edited by
J. Michael Fitzpatrick, Milan Sonka, Proceedings of SPIE Vol. 5370
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.536126

2. MATERIALS AND METHODS

2.1 Materials

The data set we used in this study contained 104 cases with 243 images. All the data were collected with institutional review board (IRB) approval. The raw mammograms in this data set were acquired with a GE FFDM system at a pixel size of $100\mu m \times 100\mu m$ and 14 bits per pixel. Most of the cases had two mammographic views, the craniocaudal (CC) view and the mediolateral oblique (MLO) view or the lateral view, except for 12 cases that had three views and 1 case with only one view. The total number of the masses in this data set is 106, of which 104 were biopsy-proven and 2 were followed up. The true locations of the masses were identified by an experienced breast radiologist.

2.2 Methods

Our CAD system consists of five processing steps: 1) preprocessing by using multi-scale enhancement, 2) pre-screening of mass candidates, 3) identification of suspicious objects, 4) extraction of feature parameters, and 5) classification between the normal and the abnormal regions by using rule-base and linear discrimination analysis (LDA) classifiers. The block diagram for the scheme is shown in Figure 1.

FFDMs generally are pre-processed with proprietary methods before being displayed to readers. The image pre-processing method used depends on the manufacturer of the FFDM system. In an effort to develop a CAD system that is less dependent on specific FFDM systems, the raw digital images are used as input to our system. A preprocessing scheme based on a multi-resolution method[20] has been developed for image enhancement. This scheme consists of three steps. First, the boundary of the breast is detected automatically by using Otsu's method[21]. Second, the Laplacian pyramid is used to decompose the image into multi-scales. A nonlinear weight function is designed to enhance each high-pass component. Finally, the Gaussian pyramid is used to reconstruct the multi-scales. The block diagram for the scheme is shown in Figure 2. An example of an original mammogram and the enhanced mammogram are shown in Figs. 3(a) and 3(b), respectively.

In our previous CAD system developed on digitized screen-film mammograms (SFM), an adaptive density-weighted contrast enhancement (DWCE) filter[18] was developed for prescreening. Although the DWCE filter using the gray level information can identify the suspicious location of masses on mammograms with high sensitivity, the prescreening objects often include a large number of enhanced normal breast structures. In this study, we investigate the use of a new method that combines gradient field information and gray level information to detect the mass candidates on the FFDMs. Gradient field information is commonly used in computer vision or other fields to extract objects or intensity field distributions. Kobatake et al[22] designed a filter, referred to as an iris filter, to calculate the convergence of gradient index around each pixel on SFMs which provided shape information for detection of masses. An extension of the iris filter, referred to as an adaptive ring filter, was developed by Wei et al[23] for detection of lung nodules on chest x-ray images. In this study, we have developed a two-stage gradient field analysis method which does not only use the shape information of masses on mammograms (an extension of the adaptive ring filter) but also incorporates the gray level information by using a region growing technique in the second stage to refine the gradient field analysis.

After prescreening, the suspicious objects are identified by using a clustering based region growing method. Figures 3(c) and 3(d) show the initial detection locations and the grown objects, respectively. For each suspicious object, eleven morphologic features are extracted and rule-based and linear classifiers are trained to remove the detected normal structures that are substantially different from breast masses. Global and local multiresolution texture analysis[24,25] are performed in each region of interest by using the spatial gray level dependence matrix. A new gradient field feature is extracted from each suspicious object and added to the feature space for false positive (FP) reduction. Finally, LDA classification is used to identify potential breast masses. Figure 3(e) shows the final detected objects, and Figure 3(f) shows the locations of these objects superimposed on the mammogram, respectively. Further details of this algorithm can be found in the literature[19].

3. RESULTS

We randomly separated the cases in our data set into two independent equal sized data sets: the training data set contained 52 cases with 120 images and the test data set contained 52 cases with 123 images, respectively. Both the mass detection system with DWCE filtering and that with the two-stage gradient field analysis were trained with the training set, the performance of the two trained systems were compared using the test data set. Our CAD system with the DWCE filter for prescreening of mass candidates achieved a case-based sensitivity of 70% and 80% at 1.4 and 1.7 FP marks/image, respectively. When the DWCE filter was replaced by the gradient field analysis for prescreening, the FP marks/image was reduced to 1.0 and 1.4 at the sensitivity of 70% and 80%, respectively. After the addition of the gradient field feature, the FP was further reduced to 0.8 and 1.3 FP marks/image, respectively, at these sensitivities. Alternatively, the new method can achieve a case-based sensitivity of 88% at 1.7 FP marks/image. Figures 4 and 5 show the comparison of performance by using image-based FROC and case-based FROC curves, respectively.

4. DISCUSSION AND CONCLUSIONS

Several FFDM systems have been approved for clinical applications. It is important to develop a CAD system that can easily be adapted to images acquired by FFDM systems from different manufacturers. In this work, we developed a CAD system that uses the raw FFDMs as the input. Our previous CAD system which was developed on digitized mammograms was adapted to FFDMs by using a new prescreening method that employed gradient field analysis and by retraining the processing parameters. A gradient field feature was extracted for further false positive reduction. The gradient field analysis in combination with the gradient field feature can reduce FPs in mass detection on FFDMs. It was found that our CAD system achieved a cased-based sensitivity of 70%, 80%, and 88% at 0.8, 1.3, and 1.7 FP marks/image, respectively. Further study is underway to improve the CAD system using a larger data set.

ACKNOWLEGMENTS

This work is supported by USPHS grant CA95153, U. S. Army Medical Research and Materiel Command grants DAMD 17-01-1-0326 and DAMD 17-02-1-0214. The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred.

REFERENCES

1. J. R. Harris, M. E. Lippman, U. Veronesi, and W. Willett, "Breast Cancer," *N Engl J Med* **327**, 319-328, 1992.
2. S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics, 1998," *CA Cancer J Clin* **48**, 6-29, 1998.
3. C. Byrne, C. R. Smart, C. Cherk, and W. H. Hartmann, "Survival advantage differences by age: evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project," *Cancer* **74**, 301-310, 1994.
4. S. A. Feig and R. E. Hendrick, "Risk, Benefit, and Controversies in Mammographic Screening," *In: Syllabus: A categorical Course in Physics Technical Aspects of Breast Imaging*, 119-135, Eds. A. G. Haus and M. J. Yaffe, Radiological Society of North America, Inc, Oak Brook, IL, 1993.
5. C. R. Smart, R. E. Hendrick, J. H. Rutledge, and R. A. Smith, "Benefit of mammography screening in women ages 40 to 49 years: current evidence from randomized controlled trials," *Cancer* **75**, 1619-1626, 1995.
6. H. Seidman, S. K. Gelb, E. Silverberg, N. LaVerda, and J. A. Lubera, "Survival experience in the Breast Cancer Detection Demonstration Project," *CA Cancer J Clin.* **37**, 258-290, 1987.
7. B. J. Hillman, L. L. Fajardo, T. B. Hunter, and e. al, "Mammogram interpretation by physician assistants," *AJR* **149**, 907-911, 1987.
8. L. W. Bassett, D. H. Bunnell, R. Jahanshahi, R. H. Gold, R. D. Arndt, and J. Linsman, "Breast cancer detection: one versus two views," *Radiology* **165**, 95-97, 1987.
9. M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," *Clinical Radiology* **44**, 13-15, 1991.

10. J. A. Harvey, L. L. Fajardo, and C. A. Innis, "Previous mammograms in patients with impalpable breast carcinomas: Retrospective vs blinded interpretation," *American Journal of Roentgenology* **161**, 1167-1172, 1993.
11. R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology* **184**, 613-617, 1992.
12. C. Beam, P. Layde, and D. Sullivan, "Variability in the interpretation of screening mammograms by US Radiologists," *Arch Intern Med* **156**, 209-213, 1996.
13. F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, "Novel technologies in breast imaging: National Cancer Institute perspective," *Society of Breast Imaging Conference*, 153-156, Orlando, Florida, 1995.
14. C. J. Vyborny, "Can computers help radiologists read mammograms?," *Radiology* **191**, 315-317, 1994.
15. H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," *Investigative Radiology* **25**, 1102-1110, 1990.
16. L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. F. O'Shaughnessy, E. A. Sickles, L. Tabar, C. J. Vyborny, and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," *Radiology* **215**, 554-562, 2000.
17. T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology* **220**, 781-786, 2001.
18. N. Petrick, H. P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Transactions on Medical Imaging* **15**, 59-67, 1996.
19. N. Petrick, H. P. Chan, B. Sahiner, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evalution of a mass detection algorithm for computer-aided diagnosis: Experience in 263 patients.," *Radiology* **224**, 217-224, 2002.
20. P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications* **COM-31**, 337-345, 1983.
21. N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. System, Man, Cybernetics* **9**, 62-66, 1979.
22. H. Kobatake and S. Hashimoto, "Convergence Index Filter for Vector Fields," *IEEE Transactions on Image Processing* **8**, 1029-1038, 1999.
23. J. Wei, Y. Hagihara, and H. Kobatake, "Detection of Rounded Opacities on Chest Radiographs Using Convergence Index Filter," *ICIAP*, 757-761, Venice, 1999.
24. D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograpms: Multiresolution texture analysis," *Medical Physics* **22**, 1501-1513, 1995.
25. D. Wei, H. P. Chan, N. Petrick, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "False-positive reduction technique for detection of masses on digital mammograms: global and local multiresolution texture analysis," *Medical Physics* **24**, 903-914, 1997.

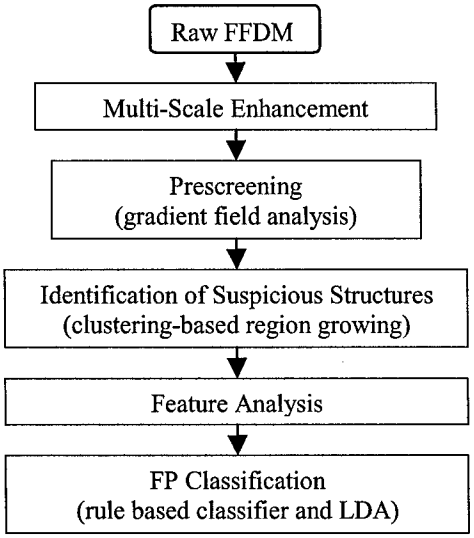


Figure 1: The block diagram of CAD algorithm for mass detection on FFDMs.

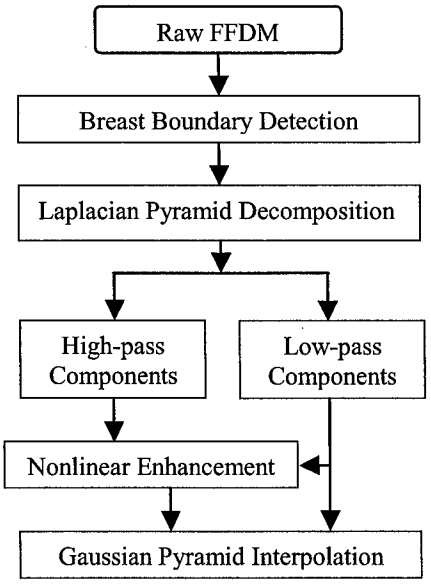
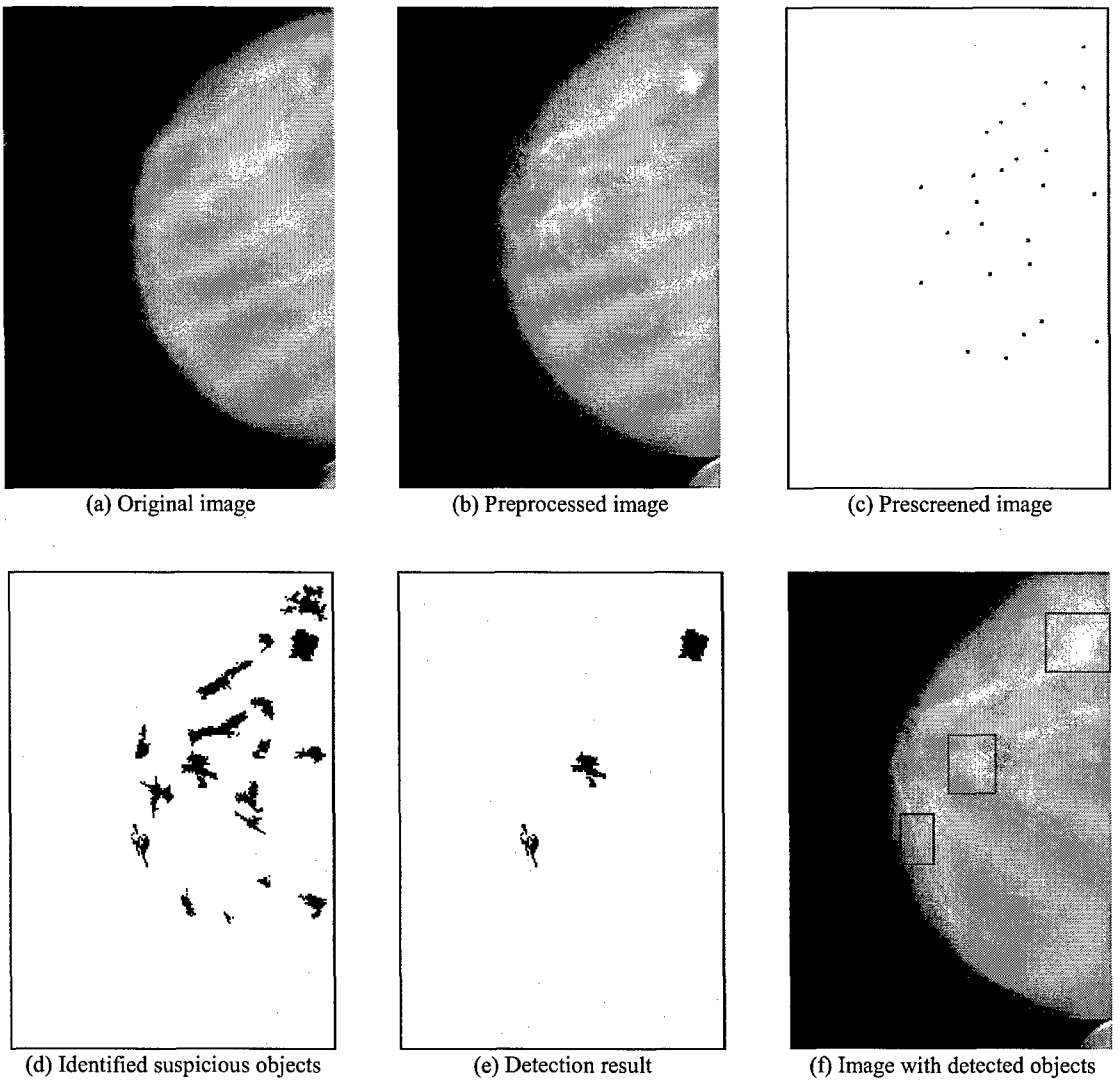


Figure 2: The block diagram for preprocessing of raw FFDM images by multiscale enhancement.



(a) Original image (b) Preprocessed image (c) Prescreened image

(d) Identified suspicious objects (e) Detection result (f) Image with detected objects

Figure 3: An example demonstrating the processing steps with our computer-aided mass detection system.

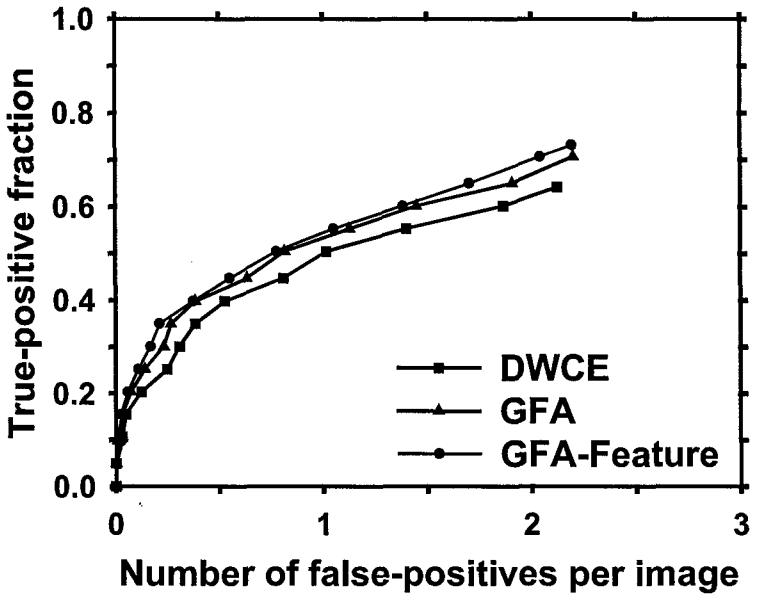


Figure 4: Image-based FROC curves. DWCE: prescreening using DWCE filter. GFA: prescreening using gradient field analysis. GFA-Feature: prescreening using gradient field analysis and the addition of the gradient field feature for FP reduction.

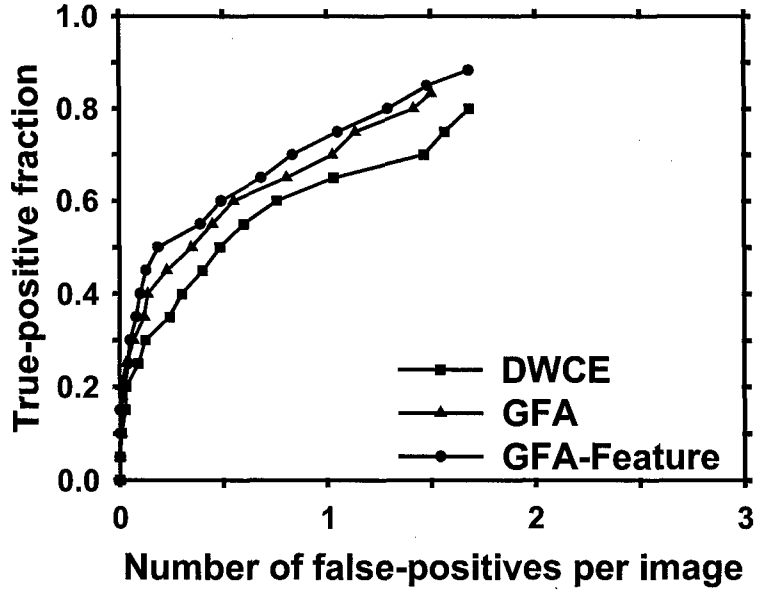


Figure 5: Case-based FROC curves. DWCE: prescreening using DWCE filter. GFA: prescreening using gradient field analysis. GFA-Feature: prescreening using gradient field analysis and the addition of the gradient field feature for FP reduction.

ROC Study of the Effects of Computer-Aided Interval Change Analysis on Radiologists' Characterization of Breast Masses in Two-View Serial Mammograms

Lubomir Hadjiiski*, Mark A. Helvie, Berkman Sahiner, Heang-Ping Chan, Marilyn A. Roubidoux, Alexis Nees, Nicholas Petrick[a], Caroline Blane, Chintana Paramagul, Janet Bailey, Stephanie Patterson, Katherine Klein, Dorit Adler, Michelle Foster, Joseph Shen

Department of Radiology, University of Michigan, Ann Arbor, MI 48109;
[a]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, MD 20857

ABSTRACT

We have previously evaluated the effects of computer-aided diagnosis (CAD) on radiologists' characterization of malignant and benign breast masses in single-view serial mammograms. In this study, we conducted observer performance experiments with ROC methodology in which the radiologists read the serial mammograms in two-views (CC and MLO) without and with CAD. 47 temporal pairs of two-view serial mammograms (27 malignant and 20 benign) containing masses were chosen from 39 patient files and digitized. The corresponding masses on each temporal pair were analyzed by the CAD program. For this data set, the computer classifier achieved a test A_z value of 0.90. Five MQSA radiologists assessed the two-view temporal pairs and provided estimates of the likelihood of malignancy without and then with CAD. For the five radiologists, the average A_z was 0.81 (range: 0.72 – 0.88) without CAD and improved to 0.88 (range: 0.86 – 0.90) with CAD. The improvement was statistically significant ($p=0.038$). In comparison, the test A_z value of the computer classifier for single view analysis was 0.87. The average A_z of the radiologists for reading the single view temporal pairs without CAD was 0.78 (range: 0.73 – 0.83) and was improved significantly ($p=0.002$) to 0.84 (range: 0.77 – 0.88) with CAD. CAD using interval change analysis can significantly improve radiologists' accuracy in classification of masses. Classification based on information from two-views is more accurate than that based on single view for both the radiologists and the computer classifier. CAD can further improve radiologists' performance even in two-view reading.

Keywords: Computer-Aided Diagnosis, Interval Changes, ROC Observer Study, Classification, Mammography, Malignancy.

1. INTRODUCTION

Mammography is currently the most sensitive method for detecting early breast cancer, and it is also the most practical screening exam[1] compared with other breast imaging techniques. However, the specificity of mammography is relatively low, only 15-30% of suspected breast lesions recommended for biopsy are actually malignant[2,3]. The unnecessary biopsies increase health care costs and cause patient anxiety and morbidity. If the specificity of differentiating malignant and benign mammographic lesions can be improved, the efficacy of mammography will be enhanced.

One of the important techniques that radiologists use in mammographic interpretation is to compare the current mammograms of a patient with those obtained in previous years, if available. The interval change information can help the detection of abnormalities, and identification of malignant breast lesions. It is shown that comparison with prior mammograms can improve both the sensitivity and specificity in breast cancer diagnosis[4,5].

* L. H. (correspondence): e-mail:lhadjiski@umich.edu

Medical Imaging 2004: Image Processing, edited by
J. Michael Fitzpatrick, Milan Sonka, Proceedings of SPIE Vol. 5370
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.535892

In an early investigation, Chan et al.[6] demonstrated that computer-aided diagnosis (CAD) could improve significantly radiologists' detection of subtle mammographic microcalcification in an ROC study, This promising result stimulated continued development of CAD systems. To date, a number of CAD algorithms have been developed to detect suspicious masses and microcalcifications and to distinguish malignant and benign lesions on mammograms. Several ROC studies have shown that CAD systems could improve radiologists' accuracy in characterization of breast lesions. It has also been reported that CAD systems can increase the detection of breast cancers on screening mammograms in clinical practice[7,8].

Chan et al[9] performed an observer study to evaluate the effects of CAD, designed for characterization of malignant and benign masses on single view mammograms[10], on radiologists' diagnostic accuracy. They found that the radiologists' accuracy for classification of masses as malignant or benign in terms of the area under the receiver operating characteristic (ROC) curve (A_z) was significantly improved ($p=0.022$ for one-view reading and 0.007 for two-view reading) with CAD compared to that without CAD. Huo et al[11] also conducted an observer study with 12 radiologists to classify masses on multiple views of mammograms. They also found that the radiologists' performance in terms of A_z was significantly improved ($p=0.001$) with computer aid. Jiang et al[12] performed an observer study to evaluate the effect of CAD on radiologists' classification of microcalcification clusters on mammograms. They found that with the computer aid the radiologists achieved a statistically significant improvement ($p<0.0001$).

The CAD systems for lesion classification so far employed information from a single exam[10,12-17]. Based on the experiences of radiologists, it can be expected that even higher accuracy may be achieved if the computer can utilize the interval change information from multiple exams for classification. We recently[18] developed a classification scheme that combines prior and current information automatically extracted from masses on prior and current mammograms, respectively. We found that the classifier using the combined prior and current information performed significantly better ($p=0.015$) in terms of A_z than the classifier using current information alone. Additionally we used the temporal classifier as a CAD system and studied its effect on radiologists' performance for characterization. Previously we studied radiologists' performance of characterizing malignant and benign masses in single-view serial mammograms with and without CAD[19,20]. We found a statistically significant improvement ($p=0.0003$) in the radiologists' performance when they used CAD compared to their performance without CAD. Since it has been reported that radiologists have higher detection and classification accuracy in interpretation of two-view mammograms[21,22], it remains to be shown that CAD can still improve radiologists' performance for characterization of masses in two-view reading. The current study investigated the effects of CAD on assisting radiologists in evaluating interval changes in two-view serial mammograms. To our knowledge, this is the first ROC experiment to evaluate the effects of a computer classifier using two-view interval change information on radiologists' diagnosis of breast cancers.

2. MATERIALS AND METHODS

2.1 Data set

We used a set of 47 two-view temporal pairs of mammograms containing biopsy-proven masses on the current mammograms from our database. The mammograms in the database were digitized consecutively from the patients who had undergone breast biopsy in our department. The selection criterion used in the current study was that the case had corresponding CC and MLO serial exams in which a corresponding mass could be identified. The mammograms thus contained masses covering a range of sizes and conspicuity that will be seen in clinical practice. The data set consisted of 168 mammograms from 39 patients. The mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of $50\mu m \times 50\mu m$ and 4096 gray levels. The image matrix size was reduced by averaging every 2 x 2 adjacent pixels and down-sampled by a factor of 2 to obtain images with a pixel size of $100\mu m \times 100\mu m$ for analysis by the computer.

There were 39 biopsy proven masses (21 malignant and 18 benign) in the 39 cases. The 168 mammograms contained corresponding CC and MLO mammographic views and multiple serial examinations of the masses including the examination when the biopsy decision was made. By matching masses on the CC and MLO views from two different examinations, a total of 47 two-view temporal pairs were formed, of which 27 were malignant and 20 benign. Since all cases in this data set had undergone biopsy, the benign masses in this set could not be distinguished easily from the malignant ones based on mammographic criteria. For the malignant masses in this data set, the average mass size was

7.9 mm on the prior mammograms and 11.2 mm on the current mammograms. The corresponding sizes were 9.5mm and 11.3 mm, respectively, for the benign masses.

To simulate a more realistic clinical situation 3 additional two-view temporal pairs containing corresponding normal structures in the serial mammograms were also included. In this way the radiologist also had to distinguish mass-mimicking fibroglandular tissue from malignant masses. The temporal pairs had a time interval of 6 to 48 months. More than 65% of the pairs had a time interval of 12 months.

2.2 Design of classifier for classification of masses in serial mammograms

We previously have developed a novel classification technique that utilizes the current and prior information on serial mammograms to characterize the masses on single-views. The classification technique has been described in detail elsewhere[18]. A brief description of the method follows. Initially a region of interest (ROI) containing the mass was defined by a radiologist on both the current and prior mammograms. An automatic segmentation of the mass within each ROI was performed based on an active contour model [23]. A set of texture, morphological, and spiculation features was extracted for each mass.

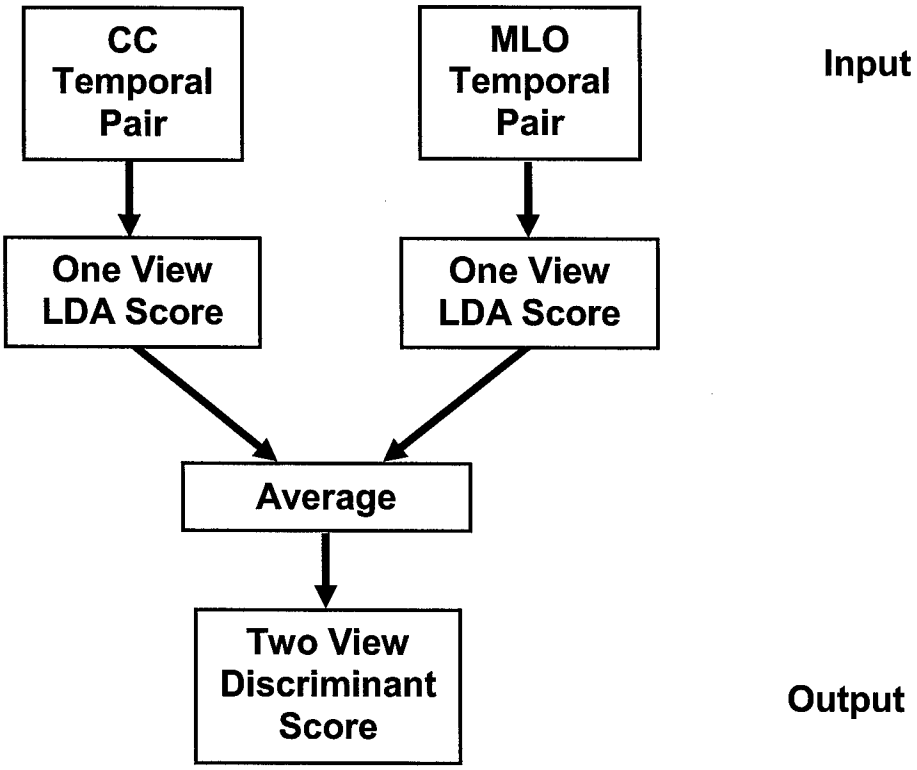


Figure 1. Block-diagram of the classification method.

The texture features were based on run-length statistics (RLS) matrices[24]. The RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST)[10]. The RBST maps a band of pixels surrounding the mass onto a rectangular region. Texture measures were extracted from the vertical and horizontal gradient images derived from the RBST image in two directions[10]. For each ROI, a total of 20 RLS features were calculated. Morphological features were extracted from the automatically segmented mass shape and gray levels[23,25]. Spiculation features were extracted by using the statistics of the image gradient direction relative to the normal direction

to the mass border in a ring of pixels surrounding the mass[23]. A total of 35 features (20 RLS, 12 morphological and 3 spiculation) were extracted from each ROI. Additionally, 35 difference features were obtained by subtracting a prior feature from the corresponding current feature.

A linear discriminant analysis (LDA) classifier was trained and tested using a "leave-one-case-out" resampling scheme. Stepwise feature selection was employed to select the most effective feature subset in each training cycle. An average of 7 features were selected for the classification task from the training subsets.

The two-view classifier was designed based on the single-view classifier. The method is summarized in the flowchart shown in Figure 1. Each two-view temporal pair consisted of the CC and MLO temporal pairs of the same mass. The single-view scores for the corresponding CC and MLO temporal pairs were averaged to produce a two-view score

A relative malignancy rating by the computer classifier on a scale of 1 to 10 was provided to the radiologists for the reading with CAD. The relative malignancy rating was obtained by linearly scaling the classifier output within the interval between 1 and 10 and then rounding the result to the closest integer. A higher rating corresponded to a higher likelihood of being malignant. Gaussian functions were fitted to the distributions of the two-view scores of the malignant and benign samples to obtain a fitted binormal distribution with the classifier's malignancy ratings scaled to the range of 1 to 10 (Figure 2). The fitted distribution was displayed on the graphical user interface as a reference when the radiologist evaluated the cases using CAD.

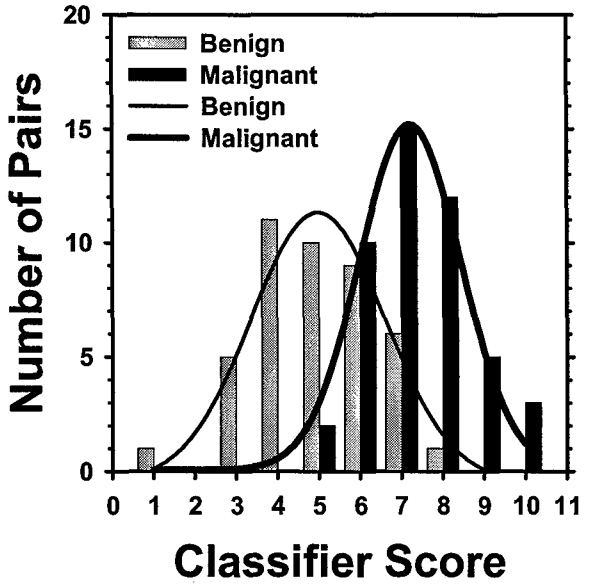


Figure 2. Binormal distribution fitted to the histogram.

2.3 Radiologist's classification of masses in two-view serial mammograms

The observer study was designed to compare radiologists' performance on the classification of malignant and benign breast masses with and without CAD on two-view temporal pairs of mass ROIs. The ROIs extracted from the current and the prior CC and MLO mammograms containing the corresponding mass were displayed side-by-side on a display monitor. The observers' performance was evaluated under two reading conditions – reading with and without CAD. In the first reading condition, the radiologist read the temporal image pair of the mass without computer aid. In the second reading condition, the radiologist read the temporal pair with computer classifier's relative malignancy rating of the mass displayed on the screen. The observer was asked to provide an estimate of the likelihood of malignancy of the mass in a 100-point rating scale under each reading condition. Four MQSA radiologists and one breast imaging fellow participated as observers in this study.

A counter-balanced design was used in arranging the reading orders in different modes and the case orders in different reading sessions for the observers. This approach would minimize the potential effects such as learning, fatigue, and memorization on the outcomes of the observer experiments. A graphical user interface was developed for the purpose of presenting the temporal pairs of mass ROIs to the radiologists and recording their ratings. Each observer underwent a training session before the actual reading sessions to familiarize them with the performance of the CAD system and the experimental procedure.

2.4 ROC analysis

The likelihood of malignancy ratings of the individual observers for the two reading conditions were analyzed by using ROC methodology. A binormal ROC curve was fitted to each observer's 100-point rating data for each reading condition by the LABROC program using maximum likelihood estimation. The classification accuracy was quantified by using the total area under the fitted ROC curve, A_z,. The slope and the intercept parameters for the individual ROC curves were also estimated by the LABROC program. For each reading condition, the average performance of the radiologists was estimated as the area under an average ROC curve, which was derived from the average slope and intercept parameters of the 5 individual observer's ROC curves for that reading condition. The statistical significance of the difference in A_z between the two reading conditions was estimated by the Student's two-tailed paired t-test on the 5 pairs of individual observer's A_z values.

3. RESULTS

The evaluation results for the five radiologists are presented in Fig 3 and Fig 4. The computer classifier achieved a test A_z value of 0.90. For the five radiologists the average A_z for the likelihood of malignancy was 0.81 (range: 0.72 – 0.88) without CAD and improved to 0.88 (range: 0.86 – 0.90) with CAD (Fig 3). The improvement was statistically significant (Student's two-tailed paired t-test, p=0.038). The average ROC curves for the 5 observers when reading with and without CAD were plotted in Fig.4. In comparison, the test A_z value of the computer classifier for single view analysis was 0.87. The average A_z of the radiologists for single view temporal pairs without CAD was 0.78 (range: 0.73 – 0.83) and was improved significantly (Student's two-tailed paired t-test, p=0.002) to 0.84 (range: 0.77 – 0.88) with CAD.

Both the malignant and benign cases in the data set were recommended for biopsy. It is possible to conclude that the data set consisted of difficult cases since the radiologists observed change in the benign masses and recommended biopsy. Each individual A_z value for the five radiologists in the evaluation mode without CAD was smaller than the computer classifier's A_z value. The relatively low accuracy of the radiologists in classifying the masses can likely be attributed to the fact that the data set was difficult.

Four of the five radiologists improved their accuracy in classification of the malignant and benign masses when the CAD system was available as a second opinion. The classification accuracy of one radiologist was not changed with the computer aid. Two-view temporal information compared to the single-view temporal information improved the radiologists' accuracy both in the reading with and without use of CAD. However the difference in A_z between the two-view and one-view readings was not statistically significant.

One radiologist achieved an A_z equal to that of the computer classifier under the reading condition with CAD. We did not observe specific differences between the breast imaging fellow compared to the MQSA radiologists. The improvement in A_z ranged between 0.06 and 0.15.

4. CONCLUSION

We reported the results of an observer ROC study that was performed to evaluate the effects of computer-aided diagnosis on radiologists' characterization of masses on serial mammograms. CAD using interval change analysis can significantly improve radiologists' accuracy in classification of masses (p=0.038). Classification based on temporal information from two views is more accurate than that based on single view for both the radiologists and the computer classifier. CAD can further improve radiologists' performance even in two-view reading, when the radiologists have more information to make a decision. These results suggest that CAD may be helpful in improving the radiologists' accuracy of characterizing malignant and benign mass and thus has a potential to reduce unnecessary biopsies.

ACKNOWLEDGMENTS

This work was supported by USAMRMC grants DAMD17-98-1-8211, DAMD17-02-1-0489, and DAMD17-02-1-0214. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC program.

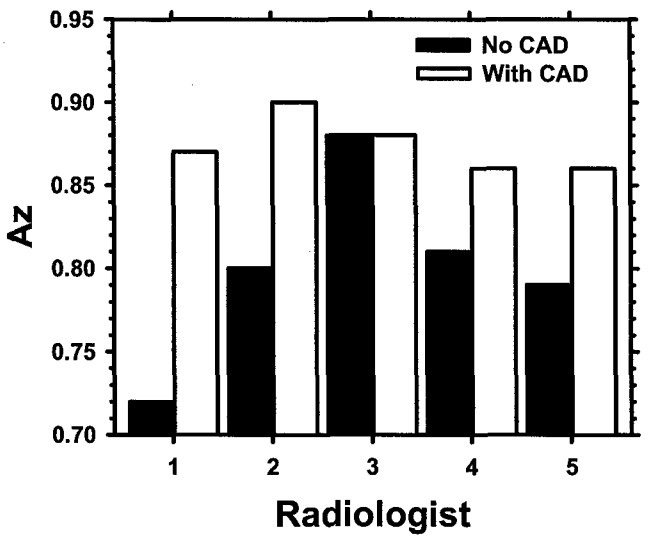


Figure 3. The area under ROC curve, A_z, for the characterization of masses in 47 two-view pairs of serial mammograms by 5 radiologists under two reading conditions: without CAD and with CAD. The average A_z for the two reading conditions: no CAD (A_z=0.81), with CAD (A_z=0.88).

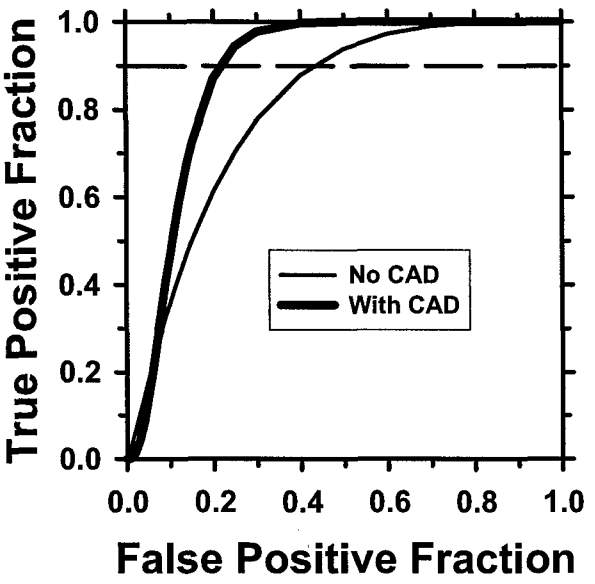


Figure 4. ROC curves for the reading conditions without CAD and with CAD by the 5 radiologists. The average area under the ROC curve for the two reading conditions: No CAD (A_z=0.81), With CAD (A_z=0.88). The difference is statistically significant (Student paired t-test, p=0.038).

REFERENCES

1. L. Tabar and P. B. Dean, "The Control of Breast Cancer through Mammography Screening," *Radiologic Clincs of North America* 25, 961, 1987.
2. D. B. Kopans, "The positive predictive value of mammography," *American Journal of Roentgenology* 158, 521-526, 1991.
3. D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Current Opinion in Radiology* 4, 123-129, 1992.
4. L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: usefullness and costs," *American Journal of Roentgenology* 163, 1083-1086, 1994.
5. E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3183 consecutive cases," *Radiology* 179, 463-468, 1991.
6. H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," *Investigative Radiology* 25, 1102-1110, 1990.
7. M. A. Helvie, L. M. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, and S. B. Lo, "A Non-Commercial CAD System for Breast Cancer Detection on Screening Mammograms Achieves High Sensitivity : A Pilot Clinical Trial," *Radiology* 225(P), 459, 2002.

8. T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology* 220, 781-786, 2001.
9. H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology* 212, 817-827, 1999.
10. B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics* 25, 516-526, 1998.
11. Z. M. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast Cancer: Effectiveness of Computer-aided Diagnosis - Observer Study with Independent Database of Mammograms," *Radiology* 224, 560-568, 2002.
12. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Academic Radiology* 6, 22-33, 1999.
13. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Leung, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," *Physics in Medicine and Biology* 42, 549-567, 1997.
14. Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Academic Radiology* 5, 155-168, 1998.
15. J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," *IEEE Transactions on Medical Imaging* 12, 664-669, 1993.
16. L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Transactions on Medical Imaging* 18, 1178-1187, 1999.
17. G. D. Tourassi, M. K. Markey, J. Y. Lo, and C. E. Floyd, "A neural network approach to breast cancer diagnosis as a constraint satisfaction problem," *Medical Physics* 28, 804-811, 2001.
18. L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. N. Gurcan, "Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses," *Medical Physics* 28, 2309-2317, 2001.
19. L. M. Hadjiiski, H. P. Chan, B. Sahiner, M. A. Helvie, M. Roubidoux, C. Blane, C. Paramagul, N. Petrick, J. Bailey, K. Klein, et al., "ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms," *Proc. SPIE Medical Imaging* 5032, 94-101, 2003.
20. L. M. Hadjiiski, H. P. Chan, B. Sahiner, M. A. Helvie, M. Roubidoux, C. Blane, C. Paramagul, N. Petrick, J. Bailey, K. Klein, et al., "Improvement of Radiologists' Characterization of Malignant and Benign Breast Masses in Serial Mammograms by Computer-Aided Diagnosis: An ROC Study," *Radiology* (in press), 2004.
21. E. Thurfjell, "Mammography screening: One versus two views and independent double reading," *Acta Radiologica* 35, 345-50, 1994.
22. R. Warren, S. Duffy, and S. Bashir, "The value of the second view in screening mammography," *British Journal of radiology* 69, 105-108, 1996.
23. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* 28, 1455-1465, 2001.
24. M. M. Galloway, "Texture classification using gray level run lengths," *Computer Graphics and Image Processing* 4, 172-179, 1975.
25. N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Medical Physics* 26, 1642-1654, 1999.